INRIA

UNITÉ DE RECHERCHE
INRIA-RENNES

Rapports de Recherche

N° 1545

Programme 5
Traitement du Signal,
Automatique et Productique

NOMBRE DE SOLUTIONS ET SATISFIABILITÉ D'UN PROBLÈME SAT ; *UNE APPROCHE ENSEMBLISTE, COMBINATOIRE ET STATISTIQUE*

Israël-César LERMAN

Octobre 1991

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P. 105
78153 Le Chesnay Cedex
France
Tél. (1) 39 63 55 11

★RR-1545★

IRISA

INSTITUT DE RECHERCHE EN INFORMATIQUE
ET SYSTÈMES ALÉATOIRES

Campus Universitaire de Beaulieu
35042 - RENNES CÉDEX
FRANCE
Téléphone 99.84.71.00
Télex UNIRISA 950 473 F
Télécopie 99 38 38 32

NOMBRE DE SOLUTIONS ET SATISFIABILITÉ D'UN PROBLÈME SAT ; Une approche ensembliste, combinatoire et statistique.

Israël César LERMAN

Publication interne n° 600
6 Septembre 1991
88 pages

RÉSUMÉ :

Dans le présent travail nous reconsidérons de façon systématique l'étude originale de J.C. Simon et O. Dubois [Simon & Dubois 1988, 1989] du problème SAT ; et ce, aussi bien dans ses aspects formels que statistiques. Nous apportons un traitement de la question à partir d'une représentation d'un type ensembliste et géométrique, en adoptant une approche combinatoire et statistique qui nous est usuelle en classification automatique. Ainsi, à partir d'une vision synthétique, nous reformulons, avec des apports nouveaux, ou bien une clarification du sens des résultats, les algorithmes ou calculs déjà exprimés dans la référence ci-dessus. Mais aussi et surtout, nous proposons de nouvelles algorithmiques et nous effectuons des calculs

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
(L.A. 227)
UNIVERSITÉ DE RENNES 1 I.N.S.A. DE RENNES

INSTITUT NATIONAL DE RECHERCHE
EN INFORMATIQUE ET EN AUTOMATIQUE
(LABORATOIRE DE RENNES)

originaux qui permettent de « voir » les limites de la réduction de la complexité qu'on peut espérer. A cette fin, qu'il s'agisse d'un système réel observé, ou bien résultant d'une génération aléatoire de clauses, nous distinguons de façon essentielle, le problème de l'évaluation du nombre de solutions, de celui de la reconnaissance de la satisfiabilité. Pour pouvoir entreprendre notre approche, nous sommes conduits à représenter une clause par un "cylindre logique ponctuel" et à faire jouer un rôle déterminant à la formule d'inclusion et d'exclusion". Les nouveaux algorithmes que nous présentons tiennent compte des caractéristiques statistiques marginales de la distribution des variables. Nous y introduisons le parallélisme et, de façon pertinente, notre approche de la classification automatique qui peut jouer un rôle important dans la réduction de la complexité d'un problème SAT. Sur chacun des deux aspects : évaluation et reconnaissance, des résultats significatifs sont obtenus, dans le cadre d'un système aléatoire de clauses, conformément à un modèle que nous avons coutume de considérer dans notre approche pour l'évaluation des liens (ici entre "cylindres ponctuels"), en classification, sous l'appellation "hypothèse d'absence de liaison". Ce modèle peut d'ailleurs avoir différentes formes, dont certaines sont ici mises à contribution. Une vision chaîne de Markov - que nous a communiquée F. Daudé (chercheur - thésard) - est également considérée pour l'un des modèles.

Mots clés. NP-complet ; Exclusivité logique et Indépendance statistique entre clauses ; Estimation statistique ; Complexité calcul ; Classification automatique.

NUMBER OF SOLUTIONS AND RECOGNITION OF SATISFIABILITY INSTANCES ; a set theoretic, combinatoric and statistical approach.

ABSTRACT :

This work is devoted to a new and systematic treatment of the different questions arised in the original study of J.C. Simon and O. Dubois [Simon & Dubois, 1988, 1989]. We consider both formal and statistical aspects with a set theoretic and geometrical representation. On the other hand we adopt the combinatoric and statistical point of view which is usual in our approach of data classification. Our synthetic formulation makes clearer and improve the algorithms or calculations, previously considered (cf. above reference). On the other hand and mainly, we propose new algorithms and new calculations which enable to « see » the limit of the complexity reduction that we can expect. For this respect, we essentially distinguish the problem of the evaluation of the number of solutions and that one of the recognition of the satisfiability instances. Two cases are studied concerning real observed and random systems of clauses. In our formalization we represent a clause by a logical and geometrical object that we call a " pinpoint cylinder ". On the other hand, the " inclusion and exclusion" formula plays an important rôle in our evaluations. The new algorithms that we present take into account the marginal statistical distributions of the variables. On the other hand, we introduce in these algorithms parallel procedures

and - in a relevant way - our approach in data classification. The latter can play an important role in complexity reduction of a SAT problem. In each of both aspects : evaluation of the number of solutions and recognition of the satisfiability, significant results are obtained in the context of the generation of a random system of clauses. The randomness is according to a model that we usually consider in our approach of data classification, for measuring associations (between "pinpoint cylinders", here) ; and that we call, "hypothesis of no relation". This model may take different forms. Some of them are used here. Relative to one of the latter forms, an interpretation in terms of a Markov chain, can also be considered (personal communication from F. Daudé : researcher preparing thesis).

Key words. NP complete ; Logical exclusivity and statistical independence between clauses ; Statistical estimation ; Computing complexity ; Clustering.

Table des matières.

III.2.2. Algorithme rapide de possible reconnaissance de la satisfiabilité.

III.2.3. Examen exhaustif du chargement cartésien du cube $\{0,1\}^n$.

IV. EVALUATION DE $\tilde{N} = 2^n - N$ ET RECONNAISSANCE DE LA SATISFIABILITÉ, DANS LE CAS D'UN MODÈLE ALEATOIRE.

IV.1. Introduction ; description des différents modèles aléatoires.

IV.2. Nombre moyen de solutions d'un problème SAT dans le cadre d'une hypothèse d'absence de liaison.

IV.2.1. Introduction.

IV.2.2. Loi de probabilité de card $(G^* \cap H^*)$.

IV.2.3. Evaluation par récurrence de $\mathcal{E}[\text{card}(E_1^* \cap E_2^* \cap \dots \cap E_h^*)]$ et calcul de $\mathcal{E}(\tilde{N}^*)$.

IV.2.4. Loi de probabilité de card $(E_1^* \cap E_2^* \cap \dots \cap E_h^*)$.

IV.2.5. Une interprétation en termes de chaîne de Markov.

IV.3. Reconnaissance de la satisfiabilité ; nombre moyen de clauses pour atteindre l'insatisfiabilité dans le cadre d'un modèle aléatoire.

IV.3.1. Introduction.

IV.3.2. Moyenne du nombre K de cylindres ponctuels permettant la couverture de l'espace total.

IV.3.3. Une formule de récurrence pour la loi de K.

IV.3.4. Se rendre compte que $k_0(r,n)$ est supérieur à $\mathcal{E}(K)$.

V. CONCLUSION.

I_ INTRODUCTION

Ce travail résulte directement de l'étude du rapport [Simon & Dubois 1988] dont il constitue la seule référence. C'est par rapport à cette dernière qu'il y a lieu de situer le point de vue et les résultats que nous apportons ici. Nous espérons que la plupart correspondent à des points nouveaux ; mais, notre spécialité n'étant pas l'étude de la complexité, nous comprenons que certains des aspects de notre analyse puissent se retrouver ailleurs. Notre propre domaine de recherche est en effet l'analyse combinatoire et statistique des données d'un type quelconque (logico - combinatoire ou symbolique et numérique). C'est à partir d'une suggestion de J. C. Simon que nous avons entrepris cette recherche qui reprend d'ailleurs avec un point de vue et un accent que nous espérons spécifiques les différents aspects du rapport précité. Il y a deux aspects généraux dans le travail à mener ; le premier est purement formel et combinatoire et le second est combinatoire et statistique (*introduction d'un modèle aléatoire de génération de clauses*) (cf. § III ci-dessous pour le premier et § IV ci-dessous pour le second). Quel que soit l'un des deux aspects généraux à considérer et étant donné un problème SAT défini par un système de clauses, nous distinguerons très clairement les deux problèmes NP suivants :

(i) évaluation exacte, approchée ou estimée du nombre de solutions ;

(ii) reconnaissance de la satisfiabilité.

Certes, une solution de nature non statistique au premier problème (i) peut fournir un algorithme pour le second (ii). Mais, il peut exister un algorithme de résolution du second problème (ii) qui ne peut être pertinent ou avoir un sens pour le premier.

C'est finalement cette distinction conceptuelle fondamentale qui nous a permis de clarifier la raison de l'écart entre la vérification expérimentale (cf. § 2 de Simon & Dubois 1988) et l'attendu théorique tel qu'il y est exprimé ; mais qui est en fait incorrect. Alors que l'évaluation du nombre moyen de solutions [cf. (i) ci-dessus] dans le contexte statistique, est correct. Nous obtenons des résultats significatifs dans ce cadre.

Qu'il s'agisse de l'aspect purement combinatoire et formel (cf. § III ci-dessous) ou de l'aspect combinatoire et statistique (cf. § IV ci-dessous), nous organisons notre rédaction autour de (i) d'abord, puis de (ii) ensuite. Notre démarche aura un caractère plus combinatoire et géométrique. En fait, nous travaillerons au niveau du cube logique $\{0,1\}^n$ où n est le nombre de variables dont chacune se trouve instanciée au moins une fois dans le système de clauses. Pour ce faire, nous associons à une même clause C

compremant r variables instanciées, son anti-clause que nous pouvons noter par $\tilde{C}$ et que nous représentons dans $\{0,1\}^n$, par ce que nous appellerons un "cylindre ponctuel d'ordre r" (cf. § II ci-dessous) dont on pourra temir compte de la structure géométrique dans les évaluations ou algorithmes. Dans ce contexte, la formule d'inclusion et d'exclusion aura un rôle important.

Nous avons déjà exprimé que le paragraphe IV est consacré à un aspect statistique qui a été introduit dans la référence précitée. Pour cet aspect, la notion d'indépendance en probabilité, notamment entre deux cylindres ponctuels respectivement associés à deux clauses aléatoires, a un sens très fortement établi que nous garderons. Dans ces conditions, il y a déjà lieu de souligner que la notion d'indépendance entre deux clauses telle qu'elle se trouve exprimée dans [Simon & Dubois 1988] correspond à la notion de disjonction ou d'exclusion – *au sens ensembliste du terme* – entre les deux cylindres ponctuels respectivement associés et représentant les deux anti-clauses. C'est au paragraphe II que nous préciserons les notions de base, ainsi que notre représentation des deux problèmes fondamentaux posés [cf. (i) et (ii) ci-dessus]. Ces deux problèmes sont traités au paragraphe III dans un cas réel et donc non aléatoire. Nous y montrerons le rôle très important de notre classification automatique pour réduire la complexité calcul. Le paragraphe V est réservé à une conclusion où nous chercherons sur la base de ce travail et de notre expérience de recherche

en classification automatique des données qualitatives, à donner notre appréciation sur la possibilité de la réduction de la complexité calcul face au traitement d'un problème NP.

II- LES NOTIONS DE BASE; POSITION DES DEUX PROBLEMES.

II.1- Cylindre ponctuel associé à une clause.

Soit $X = \{x_1, x_2, \ldots, x_n\}$ un ensemble de n variables booléennes et $\tilde{X} = \{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n\}$ l'ensemble des variables complémentées. Une <u>clause</u> est une disjonction de variables faisant partie de $X \cup \tilde{X}$; mais où, pour chaque indice i, $1 \leq i \leq n$, on a l'un des trois cas exclusifs suivants :

(i) seul x_i apparaît ;

(ii) seul $\tilde{x}_i$ apparaît ;

(iii) ni x_i, ni $\tilde{x}_i$ ne sont présents.

Ainsi, la formule représentant une clause comporte au plus n littéraux dont deux consécutifs sont séparés par le signe de disjonction $\vee$. Par exemple, en supposant n supérieur à 4, une clause d'ordre 3, C^3, peut être :

$$C^3 : x_1 \vee \tilde{x}_3 \vee x_4$$

Plus généralement, on pourra noter une clause d'ordre r, C^r, sous la forme :

$$C^r = y_{i_1} \vee y_{i_2} \vee \dots \vee y_{i_r} , \qquad (1)$$

où $r \leqslant n$, où – sans restreindre la généralité – $1 \leqslant i_1 < i_2 < \dots < i_r \leqslant n$ et où $y_{i_j} = x_{i_j}$ ou (exclusivement) $\tilde{x}_{i_j}$, $1 \leqslant j \leqslant r$.

Nous associons à C^r son opposé, l'anti-clause $\tilde{C}^r$ qui se met sous la forme :

$$\tilde{C}^r = \tilde{y}_{i_1} \wedge \tilde{y}_{i_2} \wedge \dots \wedge \tilde{y}_{i_r} , \qquad (2)$$

où $\wedge$ désigne la conjonction et où $\tilde{y}_{i_j} = \tilde{x}_{i_j}$ (resp. x_{i_j}) si $y_{i_j} = x_{i_j}$ (resp. $\tilde{x}_{i_j}$) , $1 \leqslant j \leqslant r$.

Nous identifions $\tilde{C}^r$ avec l'ensemble $E(\tilde{C}^r)$ – ou tout simplement E en cas de non ambiguïté – des points du cube $\{0,1\}^n$ qui satisfont ou vérifient la formule définie par le second membre de (2). E définit un cylindre que nous dirons "ponctuel" de $\{0,1\}^n$. En effet, dans le sous espace cartésien engendré par les composantes $i_1, i_2, \dots, i_r$, sa base est le point $(\alpha_{i_1}, \alpha_{i_2}, \dots, \alpha_{i_r})$ où $\alpha_{i_j} = 1$ (resp. 0) si $y_{i_j} = \tilde{x}_{i_j}$ (resp. x_{i_j}) , $1 \leqslant j \leqslant r$. Plus précisément, E est l'ensemble des sommets de $\{0,1\}^n$ dont la suite des composantes $(i_1, i_2, \dots, i_r)$ est $(\alpha_{i_1}, \alpha_{i_2}, \dots, \alpha_{i_r})$, qui est fixé. On peut d'ailleurs noter :

$$E = \{(i_1, i_2, \dots, i_r), (\alpha_{i_1}, \alpha_{i_2}, \dots, \alpha_{i_r})\} , \qquad (3)$$

où se trouve clairement indiqué l'emplacement du point qui est à la base du cylindre et la valeur du point.

Dans l'exemple ci-dessus de la clause d'ordre 3, en supposant $n = 8$, on a

$$E = \{(1,3,4), (1,0,1)\} \quad . \quad (4)$$

On peut ici, plus explicitement, noter ce cylindre ponctuel sous la forme suivante :

$$E = \{(1, \varepsilon, 0, 1, \varepsilon, \varepsilon, \varepsilon, \varepsilon)\} \quad , \quad (5)$$

où ε est une indéterminée booléenne susceptible de prendre l'une des deux valeurs logiques 0 ou 1.

Le volume (i.e. nombre de points) d'un cylindre ponctuel d'ordre r [cf. (3) ci-dessus] est 2^{n-r}. Ainsi, le volume du cylindre (5) ci-dessus est 2^5.

Soient E_1 et E_2 deux cylindres ponctuels de $\{0,1\}^n$ d'ordres respectifs r_1 et r_2 :

$$E_1 = \{(i_1, i_2, \ldots, i_{r_1}), (\alpha_{i_1}, \alpha_{i_2}, \ldots, \alpha_{i_{r_1}})\} \quad (6)$$

et

$$E_2 = \{(j_1, j_2, \ldots, j_{r_2}), (\beta_{j_1}, \beta_{j_2}, \ldots, \beta_{j_{r_2}})\} \quad . \quad (7)$$

On a les propriétés suivantes :

L'intersection $E_{12} = E_1 \cap E_2$ des deux cylindres E_1 et E_2, est un cylindre ponctuel.

Le cylindre intersection E_{12} est vide si et seulement si, il existe un indice i_p identique à un indice j_q $(1 \leqslant p \leqslant r_1 , 1 \leqslant q \leqslant r_2)$ pour lesquels :

$$\alpha_{i_p} = 1 - \beta_{j_q} . \qquad (8)$$

Si le cylindre intersection E_{12} n'est pas vide, désignons par $\{ h_1, h_2, \ldots, h_s \}$ l'ensemble des indices spécifiées à la fois dans E_1 et dans E_2 :

$$\{ h_1, h_2, \ldots, h_s \} = \{ i_1, i_2, \ldots, i_{r_1} \} \cap \{ j_1, j_2, \ldots, j_{r_2} \} , \qquad (9)$$

pour lesquels, on a donc :

$$\alpha_{h_u} = \beta_{h_u} \quad , 1 \leqslant u \leqslant s \ ; \qquad (10)$$

dans ces conditions, le cylindre ponctuel intersection E_{12} s'exprime comme suit :

$$E_{12} = \left\{ (k_1, k_2, \ldots, k_{r_1+r_2-s}) , (\gamma_{k_1}, \gamma_{k_2}, \ldots, \gamma_{r_1+r_2-s}) \right\} , \qquad (11)$$

où $(k_1, k_2, \ldots, k_{r_1+r_2-s})$ est la suite strictement croissante des indices de l'ensemble réunion

$$\{ i_1, i_2, \ldots, i_{r_1} \} \cup \{ j_1, j_2, \ldots, j_{r_2} \} \qquad (12)$$

et où, un même indice entier k_v , $1 \leqslant v \leqslant r_1 + r_2 - s$, est soit un h_u [cf. (9)], soit un i_p mais non un j_q [cf. (6)], soit un j_q mais non un i_p [cf. (7)]. Dans, respectivement, chacun des trois cas, on a :

$$(i)\ \gamma_{k_\nu} = \alpha_{k_\nu} = \beta_{k_\nu}$$

$$(ii)\ \gamma_{k_\nu} = \alpha_{k_\nu}$$

$$(iii)\ \gamma_{k_\nu} = \beta_{k_\nu} \quad . \qquad (13)$$

Le volume du cylindre intersection E_{12} est alors égal $2^{n-r_1-r_2+s}$:

$$vol.(E_{12}) = 2^{n-r_1-r_2+s} \quad . \qquad (14)$$

Pour représenter une clause C^r [cf. (1) ci-dessus], au niveau du cube $\{0,1\}^n$, nous avons préféré-pour des raisons de meilleure appréhension- associer à C^r son opposé, l'anti-clause $\tilde{C}^r$ [cf. (2) ci-dessus] que nous avons identifié avec un cylindre ponctuel. Mais, on peut directement considérer la représentation de C^r sous la forme de la réunion de r cylindres ponctuels d'ordre commun égal à 1. Plus précisément C^r sera identifié à l'ensemble $D(C^r)$ - ou D en cas de non ambiguïté- défini comme suit :

$$D = \bigcup \left\{ \{ i_\rho , \tilde{\alpha}_{i_\rho} \} / 1 \leqslant \rho \leqslant r_1 \right\} \qquad (15)$$

où $\{ i_\rho , \tilde{\alpha}_{i_\rho} \}$ est un cylindre ponctuel d'ordre 1, pour lequel $\tilde{\alpha}_{i_\rho} = 1 - \alpha_{i_\rho}$, où α_{i_ρ} a été ci-dessus défini, $1 \leqslant \rho \leqslant r$.

L'ensemble D est complémentaire dans $\{0,1\}^n$ de celui E ci-dessus

[cf. (3)] ; de sorte que son volume est $2^n - 2^{n-r}$. Cependant, nous travaillerons désormais avec la représentation (3) ci-dessus.

II.2. Notion d'"indépendance" entre clauses.

Selon [Simon & Dubois 1988], deux clauses C et C' sont "indépendantes" s'il n'existe pas une affectation de l'ensemble X des n variables x_i, $1 \leq i \leq n$, qui les contredise toutes les deux ; en d'autres termes, qui satisfait à la fois les deux anti-clauses respectivement associées $\tilde{C}$ et $\tilde{C}'$ [cf. (2) par rapport à (1) ci-dessus]. Dans ces conditions, les deux clauses C et C' sont "indépendantes" si et seulement si, les deux cylindres ponctuels associés $E(\tilde{C})$ et $E(\tilde{C}')$ [cf (3) ci-dessus] dans $\{0,1\}^n$ sont <u>exclusifs</u> (on peut également dire <u>disjoints</u>), au sens ensembliste du terme. On voit clairement que, pour qu'il en soit ainsi, il faut et il suffit que $E(\tilde{C})$ et $E(\tilde{C}')$ soient situés dans deux sous espaces complémentaires de $\{0,1\}^n$. C'est-à-dire, qu'il existe au moins une variable x_i instantiée de deux façons différentes dans C et dans C'. Imaginons en effet – sans aucunement restreindre la généralité – que la i-ème variable est instantiée sous la forme positive x_i dans C et négative $\tilde{x}_i$ dans C'. On a alors:

$$E(\tilde{C}) \subset \{i, 1\} \qquad (16)$$

et

$$E(\tilde{C}') \subset \{i, 0\}, \qquad (17)$$

conformément à la notation (3) ci-dessus. Plus précisément, $\{i,1\}$ et $\{i,0\}$ sont deux cylindres ponctuels d'ordre 1, qui forment une partition de $\{0,1\}^n$ en deux sous espaces complémentaires et on a :

$$\{0,1\}^n = \{i,1\} + \{i,0\} \quad , \qquad (18)$$

où la somme est ensembliste et correspond aussi à la somme directe entre sous espaces.

Dans ces conditions, deux clauses G et G' sont "dépendantes", ssi les cylindres ponctuels associés $E(\tilde{G})$ et $E(\tilde{G}')$ ont une intersection non vide :

$$E(\tilde{G}) \cap E(\tilde{G}') \neq \emptyset \quad , \qquad (19)$$

Pour qu'il en soit ainsi, il faut et il suffit que toute variable instantiée soit dans G, soit dans G', le soit sous une seule forme [positive ou (exclusivement) négative]. Pour une vision plus géométrique, on se référera à ce qui suit (6) et (7) ci-dessus.

Un cas particulier de "dépendance" logique entre clauses, correspond à la relation d'inclusion entre les deux cylindres ponctuels associés. En effet, l'implication

$$G \Longrightarrow G' \quad , \qquad (20)$$

qui exprime que tout littéral présent dans G apparaît dans G' sous

la même forme (positive ou négative), se traduit par

$$D(C) \subset D(C') \quad , \quad (21)$$

[cf. (15)]. De façon équivalente, on peut écrire au niveau des cylindres ponctuels respectivement associés à C et à C' :

$$E(\tilde{C}) \supset E(\tilde{C}') \quad . \quad (22)$$

Dans ce dernier cas, la base du cylindre C' est un point particulier de la base du cylindre C.

Nous avons déjà exprimé dans l'introduction (cf. §I) que compte tenu de la partie statistique de ce travail où la notion d'indépendance (entre cylindres ponctuels aléatoires) a un sens très précisément établi, nous ne parlerons plus pour désigner la notion d'"indépendance" entre clauses au sens de [Simone Dubois 1988] que de la notion équivalente d'exclusion entre cylindres ponctuels associés.

II.3. La formule d'inclusion et d'exclusion.

Soit Ω un ensemble fini et $\{E_j / 1 \leqslant j \leqslant k\}$ un ensemble de k parties non vides de Ω. La formule d'inclusion et d'exclusion permet de déterminer le cardinal de la réunion de tous les ensembles E_j, en fonction des cardinaux des intersections d'ordres respectifs 1 à k, des ensembles E_j. Plus précisément,

on a :

$$\operatorname{card}\Big(\bigcup_{1\leqslant j\leqslant k} E_j\Big)=\sum_{j}\operatorname{card}(E_j)-\sum_{\{j_1,j_2\}}\operatorname{card}(E_{j_1}\cap E_{j_2})+\cdots$$

$$\cdots+(-1)^{2p}\sum_{\{j_1,\ldots,j_{2p-1}\}}\operatorname{card}(E_{j_1}\cap\cdots\cap E_{j_{2p-1}})$$

$$+(-1)^{2p+1}\sum_{\{j_1,\ldots,j_{2p}\}}\operatorname{card}(E_{j_1}\cap\cdots\cap E_{j_{2p}})$$

$$+\cdots+(-1)^{k+1}\operatorname{card}(E_1\cap E_2\cap\cdots\cap E_k)\,. \qquad (23)$$

Dans cette formule, une expression de la forme $\{j_1,\ldots,j_q\}$ qui indexe une somme, signifie l'ensemble des parties de cardinal q de l'ensemble $\{1,2,\ldots,k\}$ de tous les indices entiers. Une telle somme comporte par conséquent $\binom{k}{q}=k!/[q!\,(k-q)!]$ termes. Par rapport aux termes génériques, le premier correspond à $p=0$ et le second, à $p=1$.

Il nous est nécessaire pour une meilleure appréhension du problème de rappeler une certaine démonstration de la formule (23) qui repose sur une décomposition de $\bigcup\{E_j/1\leqslant j\leqslant k\}$. Pour cela, il y a lieu d'introduire la notion de <u>classe élémentaire d'ordre h</u> (h entier inférieur ou égal à k). Une telle classe est une partie maximale qui se trouve à l'intersection d'exactement h parties E_j. Plus précisément, si $\{j_1,\ldots,j_h\}$ est un sous ensemble de cardinal h de $\{1,2,\ldots,k\}$, une telle classe se met sous la forme :

$$\mathcal{C}(j_1,\dots,j_h)=(E_{j_1}\cap\dots\cap E_{j_h})\cap(E_{j_{h+1}}\cup\dots\cup E_{j_k})^c$$

$$=(E_{j_1}\cap\dots\cap E_{j_h})\cap(E^c_{j_{h+1}}\cap\dots\cap E^c_{j_k}) \quad , \quad (24)$$

où $\{j_{h+1},\dots,j_k\}$ est – dans $\{1,2,\dots,k\}$ – le sous ensemble complémentaire de $\{j_1,\dots,j_h\}$ et où nous avons noté X^c, le sous ensemble complémentaire de X dans $\bigcup\{E_j \,/\, 1\leq j\leq k\}$.

Considérons le développement du second membre de (23) et un élément courant de la forme :

$$(-1)^{q+1}\sum_{\{j_1,\dots,j_q\}}\operatorname{card}(E_{j_1}\cap\dots\cap E_{j_q}) \quad . \quad (25)$$

$\mathcal{C}(j_1,\dots,j_h)$ se trouve comptée dans la somme $\binom{h}{q}$ fois ; Le coefficient binomial $\binom{h}{q}$ étant nul si h est strictement inférieur à q. Ainsi, dans le second membre de (23), $\mathcal{C}(j_1,\dots,j_h)$ se trouve en tout compté :

$$h-\binom{h}{2}+\binom{h}{3}+\dots+(-1)^{h+1}\binom{h}{h}=1 \quad (26)$$

et ce décompte s'arrête avec le h-ème terme du développement du second membre de (23), qui se présente sous la forme (25) avec $q=h$.

De façon plus précise, considérons le développement du second membre de (23) jusqu'au terme de la forme (25) correspondant à $q=2p$ et comparons les deux sommes partielles et consécutives, respectivement,

jusqu'à $(2p-1)$ et jusqu'à $2p$, qu'on peut d'ailleurs noter $S(2p-1)$ et $S(2p)$.

Dans ces conditions, toute classe élémentaire d'ordre $h \leqslant 2p-1$, se trouve comptée exactement une fois dans $S(2p-1)$ et dans $S(2p)$. Pour $h \geqslant 2p$, les nombres de fois, qu'une même classe élémentaire d'ordre h, se trouve respectivement comptée dans $S(2p-1)$ et dans $S(2p)$, sont :

$$\nu(2p-1,h) = h - \binom{h}{2} + \cdots + (-1)^{2p} \binom{h}{2p-1},$$

$$\nu(2p,h) = h - \binom{h}{2} + \cdots + (-1)^{2p} \binom{h}{2p-1} + (-1)^{2p+1} \binom{h}{2p}. \quad (27)$$

On démontre que

$$\nu(2p-1,h) = 1 + \binom{h-1}{2p-1}$$

et

$$\nu(2p,h) = 1 - \binom{h-1}{2p} \quad . \quad (28)$$

Ainsi, pour $h \geqslant 2p$, toute classe élémentaire d'ordre h, est comptée au moins deux fois dans $S(2p-1)$, alors qu'elle se trouve comptée au plus une fois dans $S(2p)$. Il en résulte la relation importante :

$$S(2p) \leqslant \operatorname{card}\left(\bigcup_{1 \leqslant j \leqslant k} E_j\right) \leqslant S(2p-1). \quad (29)$$

Maintenant, pour une même classe élémentaire d'ordre h ($h \geqslant 2p$), la différence entre le nombre de fois où elle est comptée dans $S(2p-1)$ et le nombre de fois où elle est comptée dans $S(2p)$ est exactement égale :

$$\nu(2p-1, h) - \nu(2p-h) = \binom{h}{2p} \quad . \quad (30)$$

On se rend compte que, pour h fixé, une telle différence s'amenuise au fur et à mesure que p augmente. Par conséquent, l'encadrement (29) est d'autant plus serré que p est grand.

D'autre part et généralement, plus h est grand, davantage on peut s'attendre à ce qu'une classe élémentaire d'ordre h, soit de cardinal faible. Cette propriété sera quantifiée dans la partie statistique, où nous faisons intervenir un modèle aléatoire.

Si les E_j sont des cylindres ponctuels de $\{0,1\}^n$, il en est de même – relativement à l'expression (25) ci-dessus – de $E_{j_1} \cap \cdots \cap E_{j_q}$. Ce dernier est vide si et seulement si, dans les q clauses respectivement représentées par $E_{j_1}, \ldots, E_{j_q}$, il existe une variable instanciée sous deux formes opposées (x_i et $\bar{x}_i$). Sinon et si t est le nombre total de variables instanciées dans l'une ou l'autre des q clauses, on a :

$$\text{card}\,(E_{j_1} \cap E_{j_2} \cap \cdots \cap E_{j_q}) = 2^{n-t} \quad . \quad (31)$$

A cet égard, on peut remarquer que le dernier terme du second

membre de (23) est égal soit à 0, soit à $(-1)^{k+1}$. En effet, dans les différentes clauses, chacune des n variables se trouve instanciée au moins une fois; sinon, le paramètre n du problème est plus grand qu'il ne faut.

Une somme telle que (25) comporte $\binom{n}{q}$ termes et toute la complexité provient de la nécessité de leur examen si aucune simplification, aucune approximation ni aucune estimation ne sont possibles.

II.4_ Position des deux problèmes.

On suppose la donnée de k clauses $\{C_j / 1 \leqslant j \leqslant k\}$ et on considère l'expression logique :

$$C_1 \wedge C_2 \wedge \ldots \wedge C_j \wedge \ldots \wedge C_k \quad . \quad (32)$$

Une <u>solution</u> est une affectation des n variables booléennes de $X = \{x_1, x_2, \ldots, x_i, \ldots, x_n\}$ (cf. § II. 1) pour laquelle l'expression (32) est à "vrai" ou <u>satisfaite</u>. Conformément à ce que nous avons précisé dans l'introduction (cf. § I), nous allons étudier dans l'ordre suivant les deux problèmes :

(i) nombre N de solutions ;

(ii) existence d'une solution.

Comme nous l'avons déjà exprimé, cet ordre - qui peut étonner

le mathématicien - se justifie ici, car une solution d'approximation de N, peut fournir un algorithme de reconnaissance de la satisfiabilité.

Conformément au paragraphe II.1., nous associons à chacune des clauses C_j, son anti-clause $\tilde{C}_j$, puis le cylindre ponctuel $E_j = E(\tilde{C}_j)$ du cube $\{0,1\}^n$, $1 \leqslant j \leqslant k$. Dans ces conditions, les deux problèmes (i) et (ii) ci-dessus deviennent :

(i) quel est le cardinal de l'union $\bigcup \{E_j / 1 \leqslant j \leqslant k\}$ des E_j ;

(ii) est-ce que $\bigcup \{E_j / 1 \leqslant j \leqslant k\}$ couvre tout l'espace $\{0,1\}^n$?.

En fait nous avons :

$$N = 2^n - \operatorname{card}\left(\bigcup_{1 \leqslant j \leqslant k} E_j\right) \qquad (33)$$

et la couverture de tout l'espace $\{0,1\}^n$ par $\bigcup \{E_j / 1 \leqslant j \leqslant k\}$ correspond à la non satisfiabilité de la formule (32).

Dans ces conditions, on comprendra que dans la suite de ce texte, nous nous exprimerons seulement en termes de cylindres ponctuels de $\{0,1\}^n$, ou bien en parties d'un ensemble fini, lorsque la structure géométrique de nos objets n'intervient pas.

Le fait que chacune des n variables soit instanciée - sous sa

forme positive ou sous sa forme négative — au moins une fois dans l'une des clauses C_j, $1 \leqslant j \leqslant k$, se traduit par le fait, que pour chaque i, $1 \leqslant i \leqslant n$, l'un au moins des deux cylindres ponctuels d'ordre 1, $\{i,1\}$ ou $\{i,0\}$ [cf. (9)] n'est pas vide ; c'est-à-dire, contient au moins un cylindre ponctuel E_j, $1 \leqslant j \leqslant k$.

Par ailleurs, on peut supposer que dans l'ensemble des cylindres ponctuels $\{E_j / 1 \leqslant j \leqslant k\}$, il n'en existe pas deux tels que l'un soit inclus dans l'autre. Sinon, on peut considérer l'algorithme de simplification suivant :

1- Ranger les cylindres par volume décroissant : $(E_{j_1}, E_{j_2}, \ldots, E_{j_k})$.

2- Comparer E_{j_1} avec la suite ordonnée des autres cylindres qui suivent. Si E_{j_h} $(h \geqslant 2)$ est inclus dans E_{j_1}, supprimer E_{j_h}.

3- Si E_{j_ℓ} est le premier cylindre non supprimé de la suite $(E_{j_2}, \ldots, E_{j_k})$. Revenir en 2, avec E_{j_ℓ} au lieu de E_{j_1}.

4- Et, ainsi de suite, jusqu'à fin.

III-EVALUATION DE N ET RECONNAISSANCE DE LA SATISFIABILITE DANS UN CAS RÉEL.

III.1- Aspect evaluation.

Conformément à la formule (33) du paragraphe II ci-dessus, nous chercherons à évaluer

$$\tilde{N} = 2^n - N = \mathrm{card}\left(\bigcup_{1 \leqslant j \leqslant k} E_j\right) \quad (1)$$

III.1.1.- Evaluation exacte à partir de la construction d'une partition de $E_1 \cup \cdots \cup E_j \cup \cdots \cup E_k$.

Un ensemble de parties, ici de $F = \bigcup \{E_j / 1 \leqslant j \leqslant k\}$, (forme une partition de F, si elles sont deux à deux disjointes et si leur union couvre F. En d'autres termes, la relation globale de mutuelle exclusion est binaire. L'algorithme proposé est récursif et tient compte de la structure géométrique d'un cylindre ponctuel. La partition qui sera obtenue sera en cylindres ponctuels.

Considérons au début d'une étape donnée de l'algorithme, l'ensemble des cylindres ponctuels en présence. Dans cet ensemble, repérons un couple de cylindres que nous noterons (E_g, E_h) dont l'intersection est non vide. Nous verrons ci-dessous comment effectuer au mieux ce repérage.

Conformément à (3) du paragraphe II ci-dessus, nous notons E_g et E_h de la façon suivante :

$$E_g = \{(i_1, i_2, \ldots, i_p, i_{p+1}, \ldots, i_q), (\alpha_{i_1}, \ldots, \alpha_{i_p}, \alpha_{i_{p+1}}, \ldots, \alpha_{i_q})\} \quad (2)$$

et

$$E_h = \{(i_1, i_2, \ldots, i_p, j_{p+1}, \ldots, j_r), (\alpha_{i_1}, \ldots, \alpha_{i_p}, \beta_{j_{p+1}}, \ldots, \beta_{j_r})\}, \quad (3)$$

où q est inférieur à r.

Par rapport à E_g, l'algorithme décompose E_h en une suite de cylindres ponctuels, mutuellement exclusifs et exclusifs de E_g, sauf le dernier qui est inclus dans E_g. Sans restreindre aucunement la généralité, supposons, pour rendre nos écritures plus claires, que i_q est strictement inférieur à j_{p+1} ; sinon, on reordonnera à chaque fois les indices dans la présentation des différents cylindres ponctuels de la décomposition de E_h. Cette décomposition s'articule autour des variables instanciées dans E_g et non instanciées dans E_h. Commençons par considérer la décomposition de E_h en deux classes, selon les deux hyperplans coordonnées relatifs à la direction i_{p+1} :

$$E_h = E_h \cap \{i_{p+1}, \tilde{\alpha}_{i_{p+1}}\} + E_h \cap \{i_{p+1}, \alpha_{i_{p+1}}\}, \quad (4)$$

où $\tilde{\alpha}_{i_{p+1}} = 1 - \alpha_{i_{p+1}}$ et où la somme est ensembliste. La décomposition (4) est en deux cylindres ponctuels d'ordre $r+1$. Le premier est disjoint de E_g, mais non le second qu'on peut d'ailleurs écrire sous la forme :

$$\{(i_1,\dots,i_r,i_{r+1},j_{r+1},\dots,j_r),(\alpha_{i_1},\dots,\alpha_{i_r},\alpha_{i_{r+1}},\beta_{j_{r+1}},\dots,\beta_{j_r})\}. \quad (5)$$

Dans ces conditions, on décompose ce dernier cylindre ponctuel en deux cylindres ponctuels d'ordre $r+2$, selon les deux hyperplans coordonnées relatifs à i_{r+2} :

$$E_h \cap \{i_{r+1},\alpha_{i_{r+1}}\} = E_h \cap \{(i_{r+1},i_{r+2}),(\alpha_{i_{r+1}},\tilde{\alpha}_{i_{r+2}})\}$$

$$+ E_h \cap \{(i_{r+1},i_{r+2}),(\alpha_{i_{r+1}},\alpha_{i_{r+2}})\}. \quad (6)$$

Désignons par E_h^1 (resp. E_h^2) le premier cylindre de la décomposition (4) [resp. (6)]. E_g, E_h^1 et E_h^2 sont mutuellement disjoints Mais, le second terme du second membre de (6) qui est disjoint avec E_h^2 et avec E_h^1, n'est pas disjoint avec E_g. On réitère le procédé de décomposition pour arriver à la formule finale

$$E_h = E_h^1 + \dots + E_h^\ell + \dots + E_h^{q-r} + F_h \quad , \quad (7)$$

où

$$E_h^\ell = E_h \cap \{(i_{r+1},\dots,i_{r+\ell}),(\alpha_{i_{r+1}},\dots,\alpha_{i_{r+\ell-1}},\tilde{\alpha}_{i_{r+\ell}})\} \quad (8)$$

$(1 \leqslant \ell \leqslant q-r)$ et où

$$F_h = E_h \cap \{(i_{r+1},\dots,i_q),(\alpha_{i_{r+1}},\dots,\alpha_{i_q})\} \quad . \quad (9)$$

F_h est inclus dans E_g ; d'autre part, E_g et les E_h^ℓ sont mutuellements disjoints. De sorte que

$$\operatorname{card}(E_g \cup E_h) = 2^{n-q} + \sum_{1 \leqslant \ell \leqslant q-r} 2^{n-r-\ell} \quad . \quad (10)$$

Le nombre de termes de la décomposition (7) ci-dessus est d'autant plus petit que l'entier $(q-p)$ est petit. Ce dernier représente le nombre de variables instanciées qui interviennent dans E_g, mais qui n'interviennent pas dans E_h. Dans ces conditions, relativement à la réduction de l'ensemble des cylindres ponctuels $\{E_j / 1 \leqslant j \leqslant k\}$, nous allons considérer une matrice carrée indéxée par l'ensemble des couples de cylindres ponctuels en présence et qui va évoluer au cours du processus de réduction, jusqu'à obtenir un ensemble de cylindres ponctuels "équivalent" mais exclusif. Au départ l'ensemble d'indéxation de cette matrice peut s'écrire $\{1, 2, \ldots, j, \ldots, k(1)\}^2$, où $k(1) = k$. La case (g,h), $1 \leqslant g,h \leqslant k(1)$, concerne le couple de cylindres ponctuels (E_g, E_h). Cette case sera chargée de la valeur 0 si l'intersection entre E_g et E_h est vide ; sinon, elle contiendra l'entier $p(g,h)$ qui représente le nombre de variables instanciées dans E_g, mais non instanciées dans E_h. $p(g,h)$ correspond à l'entier $(q-p)$ ci-dessus. La diagonale de cette matrice sera ignorée.

Une même étape de l'algorithme consiste à repérer la case (g,h) pour laquelle l'entier strictement positif $p(g,h)$ est le plus petit. Dans ces conditions, on réduira E_h par rapport à E_g, conformément au processus exprimé ci-dessus [cf. (7), (8) et (9)]. Le nouvel ensemble de cylindres ponctuels est alors :

$$\{E_1, \ldots, E_{h-1}, E_h^1, E_h^2, \ldots, E_h^{p(g,h)}, E_{h+1}, \ldots, E_{k(1)}\} \quad ; \quad (11)$$

et, en retiquetant les cylindres, on aboutit à :

$$\{E'_1, E'_2, \ldots, E'_j, \ldots, E'_{k(2)}\}, \quad (12)$$

où $k(2) = k(1) + r(g,h) - 1$. La nouvelle matrice sera indexée par $\{1, 2, \ldots, j, \ldots, k(2)\}^2$, elle sera reactualisée compte tenu de la suppression de E_h et de son remplacement par $\{E_h^1, \ldots, E_h^i, \ldots, E_h^{r(g,h)}\}$. Une intersection vide avec E_h implique naturellement une intersection vide avec chacun des E_h^i ; de sorte que l'on comprend qu'en réïtérant le processus, on arrive fatalement à une matrice remplie de zéros, qui indique que la décomposition en un système exclusif est achevée.

L'idée de base de l'algorithme de décomposition d'une clause C' par rapport à une clause C se trouve bien dans [Simon & Dubois 1988]. Nous espérons que nous donnons ci-dessus une vision géométrique plus synthétique et plus globale quant au traitement d'un ensemble de clauses. Si, dans la décomposition précédente, on aboutit à un système exclusif formé de ℓ cylindres ponctuels dont k_i sont d'ordre r_i, $1 \leqslant i \leqslant p$ et $\ell = k_1 + k_2 + \ldots + k_p$, on peut reprendre une formule exprimée dans la référence ci-dessus et écrire :

$$\text{card}\left(\bigcup_{1 \leqslant j \leqslant k} E_j\right) = \sum_{1 \leqslant i \leqslant p} k_i \, 2^{n-r_i} \quad (13)$$

III.1.2. *Evaluation exacte ou approchée à partir de la formule d'inclusion et d'exclusion.*

Nous avons pour ainsi dire tout exprimé au niveau du para-

graphe II.3. ci-dessus.

Considérons le terme de rang q du développement du second membre de (23) du paragraphe II.3 [cf. (25) § II.3]. On peut exprimer que si pour un sous ensemble particulier $\{j_1, \ldots, j_q\}$, $E_{j_1} \cap \ldots \cap E_{j_q} = \phi$, on élimine dans le terme de rang $q+u$ $\binom{k-q}{u}$ éléments à examiner. De sorte qu'on élimine en tout l'examem de

$$\binom{k-q}{1} + \binom{k-q}{2} + \ldots + \binom{k-q}{k-q} = 2^{k-q} - 1 \qquad (14)$$

éléments.

Mais, ce qui est un peu plus intéressant concerne le passage entre le terme de rang q et celui de rang $(q+1)$ [cf. (25) § II.3]. A cette fin désignons par

$$\mathcal{V}_q = \{ \{j_1, j_2, \ldots, j_q\} / E_{j_1} \cap E_{j_2} \cap \ldots \cap E_{j_q} = \phi \} \qquad (15)$$

dont ν_1 indiquera le cardinal. Désignons alors par ν_t le cardinal de l'ensemble des parties à t éléments de $\mathcal{V}_q$, tel que l'intersection de ces t éléments (qui sont des parties de cardinal q de $\{1, 2, \ldots, k\}$) soit une partie de cardinal $(q-1)$. Dans ces conditions, en vertu de la formule d'inclusion et d'exclusion, le nombre d'éléments du terme de rang $(q+1)$, qu'on peut éliminer de l'examem (car appartenant nécessairement à $\mathcal{V}_{q+1}$) est :

$$\left(\nu_1 - \nu_2 + \nu_3 + \ldots + (-1)^{t+1} \nu_t + \ldots + (-1)^{u+1} \nu_u\right)(k-q), \qquad (16)$$

où u est la valeur maximale observée de t.

Le plus intéressant reste l'application de la formule d'encadrement (29) du paragraphe II.3 qui donne une approximation d'autant plus précise que

$$S_{2p-1} - S_{2p} = \sum_{\{j_1,\dots,j_{2p}\}} \mathrm{card}(E_{j_1} \cap \dots \cap E_{j_{2p}}) \qquad (17)$$

est plus petit. C'est ce à quoi on doit "naturellement" s'attendre lorsque p augmente, comme d'ailleurs on le verra dans le cadre du modèle aléatoire qui sera envisagé au paragraphe IV. Autrement, nous avons fait remarquer que dans le contexte de notre problème, le dernier terme du développement (23) (§ II.3.) est en valeur absolue, au plus égal à 1.

III.1.3. *Simplification de la complexité par usage de la classification automatique.*

Nous nous situons dans le contexte combinatoire et statistique de notre approche en classification hiérarchique [Lerman 1981, 1991]. Cette dernière permet la prise en compte de n'importe quel type mathématico-logique de données, dans le cadre de principes dont la pertinence et la validité ont été largement établies, aussi bien sur le plan théorique que celui, expérimental. Elle se trouve donc particulièrement adaptée à la classification par proximité d'un ensemble $\mathcal{C} = \{E_j / 1 \leq j \leq k\}$ de cylindres ponctuels. Mais, à quoi sert une telle

classification ?

Ayant élaboré de façon adéquate un indice de proximité entre cylindres ponctuels, une forme particulière du critère d'aggrégation entre parties disjointes de G nous permettra - si elle existe - de détecter une partition de l'ensemble G des cylindres, telle que deux cylindres de deux classes distinctes, ont une intersection vide. Désignons par $\{ F_g / 1 \leqslant g \leqslant h \}$ une telle partition, on peut noter

$$F_g = \{ E^g_1, \ldots, E^g_i, \ldots, E^g_{k(g)} \} \qquad (18)$$

la g-ème classe de cardinal $k(g)$, où $k = k(1) + \ldots + k(h)$. La partition recherchée est telle que :

$$[\forall (1 \leqslant g < g' \leqslant h), \forall (1 \leqslant i \leqslant k(g), 1 \leqslant i' \leqslant k(g'))], \; E^g_i \cap E^{g'}_{i'} = \emptyset \; . \qquad (19)$$

Il est clair dans ces conditions que la complexité maximale se trouve réduite à

$$2^{\max \{ k(g) / 1 \leqslant g \leqslant h \}} \; ; \qquad (20)$$

mais, $\max \{ k(g) / 1 \leqslant g \leqslant h \}$ peut être égal à k.

Un autre intérêt tout à fait majeur de la classification, cette fois ci au moyen du critère de la vraisemblance du lien maximal [Lerman 1981, 1991], concerne le traitement par la classification de chacun des F_g, $1 \leqslant g \leqslant h$, afin de former à l'intérieur de chacun des F_g des classes de dépendance. Désignons par $\{ H_g(t) / 1 \leqslant t \leqslant u \}$ des classes "naturelles" de dépendance découvertes à l'intérieur de F_g et

formant une partition de F_g. En appliquant la formule d'inclusion et d'exclusion de façon tronquée, mais avec l'ensemble des u arguments $\{H_g(t) / 1 \leqslant t \leqslant u\}$, on aura une estimation d'autant plus précise de

$$\text{card}\left(\bigcup \{E_i^g / 1 \leqslant i \leqslant k(g)\}\right) ; \qquad (21)$$

ce qui bien sûr suppose l'application totale de la formule d'inclusion et d'exclusion au niveau de l'union des ensembles E_i^g rentrant dans la composition d'un même $H_g(t)$, $1 \leqslant t \leqslant u$.

D'autre part, la tendance naturelle du critère de la vraisemblance du lien maximal à former des classes de tailles équilibrées est un facteur réducteur de la complexité.

Nous allons maintenant présenter deux indices de proximité entre cylindres ponctuels qui sont conformes à notre approche. Le premier est assez général et suffit à notre analyse. Mais, le second est plus précis car il serre de plus près notre démarche en tenant compte de la structure propre d'un cylindre ponctuel.

Un même cylindre ponctuel définit une partie de l'ensemble plein $\{0,1\}^n$ et $\{E_j / 1 \leqslant j \leqslant k\}$ détermine un ensemble de parties de ce dernier ensemble. Relativement à la comparaison de deux parties E_g et E_h, $1 \leqslant g < h \leqslant k$, nous considérons un indice "brut" de proximité $S(g,h)$ qui représente le cardinal de l'intersection entre E_g et E_h :

$$s(g,h) = \text{card}\,(E_g \cap E_h) \quad . \qquad (22)$$

Ce dernier indice $s(g,h)$ est statistiquement normalisé par rapport à une hypothèse d'absence de liaison (h.a.l.) où au couple (E_g, E_h) de parties, on associe un couple (X,Y) de sous ensembles indépendants en probabilité, d'un ensemble Ω, où X (resp. Y) respecte par rapport à Ω, la cardinalité de E_g (resp. E_h) par rapport à l'ensemble plein, dont il est une partie. Il y a trois formes fondamentales du modèle aléatoire de l'hypothèse d'absence de liaison. Celle qui mène aux expressions calcul les plus simples et que nous avons largement adoptée est celle que nous appelons Poissonienne. Car cette forme conduit à une loi de Poisson de l'indice brut aléatoire:

$$S(g,h) = \text{card}\,(X \cap Y) \quad . \qquad (23)$$

Nous nous limiterons ici à cette forme de l'h.a.l. pour laquelle, l'espérance mathématique $\mathcal{E}[S(g,h)]$ et la variance $\text{var}[S(g,h)]$ sont toutes les deux égales à

$$\text{card}(E_g) \times \text{card}(E_h) \,/\, 2^n \quad ; \qquad (24)$$

et, si on se réfère aux expressions (1) et (2) pour E_g et E_h, on a :

$$s(g,h) = 2^{n-q-r+t} \quad . \qquad (25)$$

D'autre part:

$$\mathcal{E}[S(g,h)] = \text{var}[S(g,h)] = 2^{n-q-r} \quad . \qquad (26)$$

Le coefficient ou indice, statistiquement normalisé, se met sous la forme :

$$Q(g,h) = \frac{s(g,h) - \mathcal{E}[S(g,h)]}{\sqrt{var[S(g,h)]}} \quad . \qquad (27)$$

Nous avons ici :

$$Q(g,h) = \sqrt{2^{n-q-r}} \quad (2^r - 1) \quad . \qquad (28)$$

Pour q et r fixés, cet indice est – comme on devait s'y attendre – d'autant plus grand que r est grand ; c'est-à-dire, puisque nous avons supposé q inférieur à r, que le cylindre ponctuel E_g se trouve davantage inclus dans celui E_h. Maintenant, de façon globale, l'indice $Q(g,h)$ est d'autant plus grand ; que d'une part, $(q+r)$ est petit et que, d'autre part, r est grand. C'est-à-dire, que E_g et E_h remplissent au mieux l'espace $\{0,1\}^n$; mais que, l'apport complémentaire du chargement de E_h par rapport à celui de E_g, est plus petit. L'indice $Q(g,h)$ qui peut aussi se mettre sous la forme :

$$Q(g,h) = 2^{\frac{n}{2}} \; 2^{-\frac{1}{2}(q+r)} \; (2^r - 1) \quad , \qquad (29)$$

pondère d'une certaine façon, la croissance par rapport à p et la décroissance par rapport à $(q+r)$.

Le problème de complexité aurait été du même ordre ; avec d'ailleurs et *a priori* moins de possibilités de simplification, si $\mathcal{G} = \{E_1, \ldots, E_j, \ldots, E_k\}$, au lieu d'être formé d'un ensemble de cylindres

ponctuels de l'espace $\{0,1\}^n$, était formé d'un ensemble de parties de structure libre (c'est-à-dire, n'ayant aucune structure particulière) d'un ensemble U de cardinal de l'ordre de 2^n; l'ensemble U pouvant d'ailleurs être l'espace $\{0,1\}^n$. Nous mentionnerons d'ailleurs au paragraphe IV ce cadre.

Cependant on peut vouloir dans l'élaboration de l'hypothèse d'absence de liaison tenir compte de la structure particulière de cylindre ponctuel de E_g et de E_h dans l'espace $\{0,1\}^n$. Dans ces conditions (X,Y) [cf. ci-dessus] sera un couple de cylindres ponctuels aléatoires et indépendants de $\Omega=\{0,1\}^n$. La probabilité qu'une composante de cet espace soit spécifiée (ou instanciée) dans X (resp. Y) est $x = q/n$ (resp. $\rho = r/n$). D'autre part, q (resp. r) étant généralement petit devant n, le nombre L (resp. M) de composantes instanciées pour X (resp. Y) est parfaitement approximable par une variable aléatoire de Poisson de paramètre q (resp. r). Pour $L=\ell$ (resp. $M=m$), les 2^ℓ (resp. 2^m) instanciations pour les ℓ (resp. m) composantes, sont considérées comme également probables. Dans ces conditions, on peut déterminer la loi de probabilité de la variable aléatoire que nous désignerons ici par :

$$T(g,h) = \text{card}\,(X \cap Y) \quad . \quad (30)$$

Nous préciserons cette loi, mais dans un contexte très différent, au paragraphe IV ci-dessous, où nous verrons que :

$$\mathcal{E}[T(g,h)] = 2^{n-(q+r)} \qquad (31)$$

et

$$\mathrm{var}[T(g,h)] = 2^{2[n-(q+r)]}\left(e^{n\times p} - 1\right) . \qquad (32)$$

Il est tout à fait intéressant de constater que $\mathcal{E}[T(g,h)]$ est identique à $\mathcal{E}[S(g,h)]$; mais que, $\mathrm{var}[T(g,h)]$ est assez différent de $\mathrm{var}[S(g,h)]$ [cf. (26) ci-dessus]. L'indice, statistiquement normalisé s'écrit ici :

$$R(g,h) = \frac{s(g,h) - \mathcal{E}[T(g,h)]}{\sqrt{\mathrm{var}[T(g,h)]}}$$

$$= \frac{2^{r} - 1}{\sqrt{e^{n\times p} - 1}} . \qquad (33)$$

Ici encore, on a le même phénomène de croissance par rapport à p et de décroissance par rapport à q et à r ; mais cette décroissance est autrement pondérée dans le cas de (33) que dans le cas de (29).

L'expression (25) de l'indice brut suppose que les cylindres ponctuels E_g et E_h ont une intersection non vide. Dans le cas où E_g et E_h sont exclusifs, respectivement, les deux coefficients $Q(g,h)$ et $R(g,h)$ prennent les valeurs suivantes :

$$Q_0(g,h) = -\sqrt{2^{n-q-r}} \qquad (34)$$

et

$$R_0(g,h) = -1/\sqrt{e^{n\times p} - 1} , \qquad (35)$$

qui sont strictement négatives. Au contraire, ces coefficients prennent

des valeurs positives ou nulles en cas de non exclusivité. La valeur 0 qui traduit que card $(E_g \cap E_h)$ est à sa valeur moyenne en cas d'indépendance en probabilité, correspond exactement à $p = 0$.

A partir de là si $\mathbb{C}$ et $\mathbb{D}$ sont deux parties disjointes de l'ensemble $G = \{ E_j / 1 \leqslant j \leqslant K \}$ des cylindres ponctuels :

$$\mathbb{C} = \{ E_{j_\ell} / 1 \leqslant \ell \leqslant c \} \qquad (36)$$

et

$$\mathbb{D} = \{ E_{j_m} / 1 \leqslant m \leqslant d \} \quad , \quad (37)$$

où c (resp. d) est le cardinal de $\mathbb{C}$ (resp. $\mathbb{D}$), considérons l'indice "diamètre" d'association entre les deux sous ensembles disjoints $\mathbb{C}$ et $\mathbb{D}$:

$$\pi(\mathbb{C}, \mathbb{D}) = \min \{ S(E_{j_\ell}, E_{j_m}) / 1 \leqslant \ell \leqslant c, 1 \leqslant m \leqslant d \} \quad . (38)$$

L'indice S peut être celui Q [cf. (28)] ou celui R [cf. (33)].

Dans ces conditions, considérons la formation ascendante hiérarchique de l'arbre des classifications [Lerman 1991], conformément à l'indice $\pi(\mathbb{C}, \mathbb{D})$. Arrêtons la formation de l'arbre dès que cet indice cesse d'être positif ou nul. On aboutit alors nécessairement à la partition annoncée dont les classes F_g [cf. (18)], $1 \leqslant g \leqslant h$, remplissent la condition (19).

Relativement à l'évaluation de (21) ci-dessus après décomposition en classes de dépendance, nous verrons de façon implicite au paragraphe IV ci-dessous, pourquoi l'indépendance entre cylindres ponctuels conduit à une "bonne" approximation, au moyen de la formule tronquée d'inclusion et d'exclusion.

III.2 - Aspect reconnaissance.

III.2.1. Usage de la formule d'inclusion et d'exclusion.

Il s'agit de l'application directe de la formule d'encadrement (29) du paragraphe II.3. Pour plus de précision, nous allons distinguer dans la formule (23) (§ II.3), le cas où $k = 2q$ est pair et le cas où $k = 2q+1$ est impair.

Pour $k = 2q$, le dernier terme de (23) est négatif ou nul et correspond à $p = q$. Dans ces conditions, pour $p = 1, 2, \ldots, q$:

Si $S(2p-1) < 2^n$; alors la satisfiabilité est assurée,

Si $S(2p) \geqslant 2^n$; alors la satisfiabilité est impossible,

Si $S(2p) < 2^n \leqslant S(2p-1)$; la décision est impossible, faire alors $p = p+1$.

Pour $k = 2q+1$, le dernier terme de (23) (§ II.3) est positif ou nul et correspond à $p = q+1$. Le processus de décision correspond exactement à ci-dessus. Il aboutit nécessairement lorsque p atteind sa plus grande

valeur ; mais a de "bonnes chances" de permettre dans les cas les plus courants, une conclusion dès les premiers pas.

III.2.2. Algorithme "rapide" de possible reconnaissance de la satisfiabilité.

Nous allons considérer un algorithme dont la complexité maximale est majorée par kn^2, qui peut — en cas de conclusion positive — conduire à la reconnaissance de la satisfiabilité. L'idée générale de cet algorithme est de pouvoir reconnaître un cylindre ponctuel de $\{0,1\}^n$ qui ne peut être comblé par la réunion de tous les cylindres ponctuels associés aux différentes clauses [cf. § II.1]. La méthode consiste à entraîner par blocs l'ensemble des cylindres ponctuels associés aux clauses dans la réunion de sous espaces coordonnées (i.e. cylindres ponctuels d'ordre 1) de $\{0,1\}^n$, dont la partie complémentaire n'est pas vide.

Description de l'algorithme.

Relativement à l'ensemble des cylindres ponctuels $\{E_j / 1 \leqslant j \leqslant k\}$, désignons par $\{i_1, \alpha_{i_1}\}$ un hyperplan coordonnée, cylindre ponctuel d'ordre 1, qui contient un nombre maximal de E_j. Si $\{i_1, \alpha_{i_1}\}$ contient tous les E_j, la conclusion de satisfiabilité est acquise, puisque aucun point de $\{i_1, \tilde{\alpha}_{i_1}\}$ n'est atteint par $\cup\{E_j / 1 \leqslant j \leqslant k\}$ (qui est disjoint de $\{i_1, \tilde{\alpha}_{i_1}\}$).

Si non, désignons par $\{E_{j_1}, \ldots, E_{j_{k(1)}}\}$ le sous ensemble propre de cylindres ponctuels inclus dans $\{i_1, \alpha_{i_1}\}$ et désignons par $\{i_2, \alpha_{i_2}\}$ où $i_2 \neq i_1$, le cylindre ponctuel d'ordre 1 qui inclut un nombre maximal de cylindres parmi $\{E_j / j \notin \{j_1, \ldots, j_{k(1)}\}\}$. Ces derniers sont des cylindres dont la composante i_1 est nécessairement soit $\tilde{\alpha}_{i_1}$, soit indéterminée que nous noterons ε. Si l'ensemble de ces derniers cylindres est inclus dans $\{i_2, \alpha_{i_2}\}$, la condition de satisfiabilité est acquise. En effet, aucun point du cylindre d'ordre 2 $\{(i_1, i_2), (\tilde{\alpha}_{i_1}, \tilde{\alpha}_{i_2})\}$ ne se trouve atteint par $\cup\{E_j / 1 \leqslant j \leqslant k\}$ qui est disjoint du dernier cylindre ponctuel.

Si non, désignons par $\{E_{j_{k(1)+1}}, \ldots, E_{j_{k(2)}}\}$ le sous ensemble propre de $\{E_j / j \notin \{j_1, \ldots, j_{k(1)}\}\}$ qui se trouvent chacun inclus dans $\{i_2, \alpha_{i_2}\}$. On a alors :

$$\cup\{E_{j_i} / 1 \leqslant i \leqslant k(2)\} \subset \{i_1, \alpha_{i_1}\} \cup \{i_2, \alpha_{i_2}\} \quad . \quad (39)$$

Considérons alors le cylindre ponctuel $\{i_3, \alpha_{i_3}\}$ d'ordre 1, où i_3 n'appartient pas à $\{i_1, i_2\}$, qui contient un nombre maximal de cylindres parmi $\{E_j / j \notin \{j_1, \ldots, j_{k(2)}\}\}$, ces derniers sont des cylindres dont la composante i_1 (resp. i_2) est nécessairement soit $\tilde{\alpha}_{i_1}$, soit ε (resp. soit $\tilde{\alpha}_{i_2}$, soit ε). Et, ainsi de suite, jusqu'à une étape ℓ, où on peut aboutir à

$$\cup\{E_{j_i} / 1 \leqslant i \leqslant k(\ell)\} \subset \{i_1, \alpha_{i_1}\} \cup \ldots \cup \{i_\ell, \alpha_{i_\ell}\} \quad , \quad (40)$$

avec $k(\ell) = k$, où alors la satisfiabilité sera assurée, puisque le premier membre de (40) sera disjoint du cylindre ponctuel $\{(i_1, \ldots, i_\ell), (\tilde{\alpha}_{i_1}, \ldots, \tilde{\alpha}_{i_\ell})\}$.

Mais, on peut aussi aboutir à (40) avec $k(\ell) < k$ et avec l'impossibilité de trouver un cylindre ponctuel de la forme $\{i_{\ell+1}, \alpha_{i_{\ell+1}}\}$ où $i_{\ell+1} \notin \{i_1, \ldots, i_\ell\}$, qui puisse entraîner en son sein un des cylindres de

$$\{E_j / j \notin \{j_1, \ldots, j_{k(\ell)}\}\} \quad . \quad (41)$$

Dans ces conditions, aucune des variables X_i pour $i \notin \{i_1, \ldots, i_\ell\}$ ne se trouve instanciée dans l'un quelconque des E_j de (41). D'autre part, chacun des E_j de (41) est nécessairement de la forme :

$$\{(i_1, i_2, \ldots, i_h), (\tilde{\alpha}_{i_1}, \tilde{\alpha}_{i_2}, \ldots, \tilde{\alpha}_{i_h})\} \quad , \quad (42)$$

où $\{i_1, i_2, \ldots, i_h\} \subset \{i_1, i_2, \ldots, i_\ell\}$.

Mais alors, l'ensemble complémentaire dans $\{0,1\}^n$, du second membre de (40) et qui est le cylindre ponctuel :

$$\{(i_1, i_2, \ldots, i_\ell), (\tilde{\alpha}_{i_1}, \tilde{\alpha}_{i_2}, \ldots, \tilde{\alpha}_{i_\ell})\} \qquad (43)$$

est contenu dans chacun des cylindres ponctuels E_j de (41) dont la forme est (42) ci-dessus.

En conséquence et dans cette situation, la condition nécessaire et suffisante de satisfiabilité est que l'inclusion (40) soit stricte. On se retrouve ainsi devant un problème de satisfiabilité où k est remplacé par $k(\ell) < k$ et où, l'espace à remplir est une réunion de ℓ cylindres ponctuels d'ordre 1.

Mais, en vérité, il est très important de noter qu'on peut substituer à ce problème ℓ problèmes SAT de la forme : est-ce que l'inclusion suivante est stricte ?

$$\bigcup \{ E_i \,/\, \gamma_{k(h-1)+1} \leqslant i \leqslant \gamma_{k(h)} \} \subset \{ i_h , \alpha_{i_h} \} \quad , \quad (44)$$

pour $h = 1, 2, \ldots, \ell$; où $k(0) = 0$. Etant entendu que la satisfiabilité à l'un seulement des problèmes (44), assure la satisfiabilité de l'ensemble du problème posé !

Les ℓ problèmes pouvant être traités en parallèle, on se ramène avec un problème SAT, avec, au lieu de k,

$$\max \{ k(h) - k(h-1) \,/\, 1 \leqslant h \leqslant \ell \} \quad , \quad (45)$$

pour ce qui concerne la complexité maximale.

On appliquera donc en parallèle à chacun des problèmes (44), le processus depuis son début (cf. ci-dessus) ou d'ailleurs lorsque c'est envisageable III.2.1. ou III.2.3. ci-dessous, en arrêtant dès lors qu'une ou l'autre des relations (44) se trouve satisfaite. Sinon, on poursuit de proche en proche et de façon récursive par démultiplication, mais, en utilisant à chaque fois le traitement parallèle. Fatalement, on arrivera à la conclusion.

III.2.3. Examen exhaustif du chargement cartésien du cube $\{0,1\}^n$.

Représentons un même cylindre ponctuel E_j, associé à une

clause C_j , $1 \leq j \leq k$, [cf. (5) § II.1.], par un vecteur à n composantes, dont la i-ème , $1 \leq i \leq n$, est égale à 1, 0 ou ε, selon que dans C_j, X_i se trouve instanciée à $1, 0$ ou est indéterminée.

Relativement à l'ensemble des k vecteurs ainsi construits et à une même variable X_i, introduisons les nombres entiers suivants:

$k_i(1)$: nombre de vecteurs dont la i-ème composante est égale à 1;

$k_i(0)$: nombre de vecteurs dont la i-ème composante est égale à 0;

$k_i(\varepsilon)$: nombre de vecteurs dont la i-ème composante est égale à ε.

On a :

$$k_i(1) + k_i(0) + k_i(\varepsilon) = k, \qquad (46)$$

pour tout $i = 1, 2, \ldots, n$.

Associons à la variable X_i , $1 \leq i \leq n$, le nombre entier:

$$\ell_i = \max [k_i(1), k_i(0)].$$

On ne restreint en rien la généralité si on suppose

$$\ell_1 \geqslant \ell_2 \geqslant \cdots \geqslant \ell_i \geqslant \cdots \geqslant \ell_n ; \qquad (47)$$

c'est-à-dire, que la suite des étiquettes des variables est ordonnée selon les valeurs décroissantes de ℓ_i , $1 \leq i \leq n$. Disons tout de suite que nous effectuons ce choix pour que *a priori* dans l'algorithme qui suit, on ait "le plus de chances" de conclure à la satisfiabilité, "le plus tôt possible".

Considérons alors la première composante des k vecteurs. Si $\ell_1 = k$, la satisfiabilité est assurée. En effet, si $\ell_1 = k = k_1(1)$ [resp. $k_1(0)$], aucun

des sommets du cylindre ponctuel $\{1,0\}$ (resp. $\{1,1\}$) ne peut être touché par la réunion des cylindres ponctuels E_j, $1 \leq j \leq k$, qui est incluse dans le cylindre ponctuel $\{1,1\}$ (resp. $\{1,0\}$) (cf. notations encore adoptées au paragraphe III.2.2. ci-dessus)

Si ℓ_1 est strictement inférieur à k, construisons une table à 2 lignes et k colonnes, dont la première (resp. seconde) ligne est étiquetée par la valeur possible 0 (resp. 1) de X_1. Une telle table donne les numéros ou adresses des vecteurs décrits ci-dessus et associés aux cylindres ponctuels E_j, $1 \leq j \leq k$. Ainsi, si le j-ème vecteur a sa première composante égale à 0 ou à ε (resp. à 1 ou à ε) un booléen égal à 1 sera installé dans la j-ème colonne de la première (resp. seconde) ligne, ce booléen sera égal à 0 si cette première composante est égale à 1 (resp. 0).

Nous allons à présent examiner la deuxième composante X_2 compte tenu des valeurs possibles de la première composante X_1. On partitionne l'ensemble des adresses indiquées dans la première (resp. seconde) ligne du tableau ci-dessus où X_1 peut être égal à 0 (resp. 1) en deux classes. La première (resp. seconde) est celle où X_2 peut être égal à 0 (resp. 1) ; c'est-à dire, $X_2 = 0$ ou ε (resp. $X_2 = 1$ ou ε). Si une des configurations de (X_1, X_2), à savoir $(0,0)$, $(0,1)$, $(1,0)$ ou $(1,1)$, n'est pas possible pour l'ensemble des k vecteurs ; alors, la satisfiabilité est assurée. En effet, le cylindre ponctuel basé sur une telle configuration est disjoint de la réunion des cylindres ponctuels E_j, $1 \leq j \leq k$.

Sinon, on en déduit un tableau à 2^2 lignes et k colonnes. Les 2^2 lignes sont étiquetées par les valeurs possibles de (X_1, X_2) qui, respectivement,

sont (0,0), (0,1), (1,0) et (1,1).

Et, ainsi de suite, après l'examen de la r-ème composante, où il s'avère que toutes les configurations de $(X_1, \ldots, X_i, \ldots, X_r)$ sont possibles, on peut - conceptuellement parlant - établir un tableau à 2^r lignes et k colonnes. La l-ème ligne, $1 \leqslant l \leqslant 2^r$, est étiquetée par le l-ème - *par ordre lexicographique croissant* - vecteur à composantes 0 ou 1. Cette ligne indique les adresses, numéros des vecteurs à n composantes 0, 1 ou ε, qui représentent les différents cylindres ponctuels E_j, $1 \leqslant j \leqslant k$, (cf. ci-dessus), pour lesquels la configuration concernée est possible.

On considère alors la variable X_{r+1}. Cette dernière partitionne l'ensemble des adresses indiquées dans la l-ème ligne, en deux classes. La première concerne les vecteurs où X_{r+1} est susceptible de prendre la valeur 0 ($X_{r+1} = 0$ ou ε) et la seconde, concerne les vecteurs où X_{r+1} peut prendre la valeur 1 ($X_{r+1} = 1$ ou ε). S'il existe l, $1 \leqslant l \leqslant 2^r$, tel que l'une des deux classes est vide ; alors, la condition de satisfiabilité est assurée, car on peut exhiber un cylindre ponctuel d'intersection vide avec $\cup \{ E_j / 1 \leqslant j \leqslant k \}$. Sinon, on va être conduit - au moins conceptuellement - à la construction d'une table à 2^{r+1} lignes et k colonnes, qui va remplacer celle ci-dessus à 2^r lignes et k colonnes.

Fatalement, on va aboutir à la décision. Un tel algorithme est intéressant si, pour un temps de calcul fixé, on arrive à conclure à la satisfiabilité. Parce que, la non satisfiabilité n'est reconnaissable qu'après l'examen total de la suite de toutes les composantes : $i = 1, 2, \ldots, n$.

Pour r fixé, la complexité spaciale de la table fournissant les adresses est $k \times 2^r$. On peut substituer à cette table, de façon équivalente, deux

tables à une entrée (ou vecteurs) ; la première, donne pour la suite ordonnée des 2^r vecteurs de booléems [valeurs possibles de $(X_1, \ldots, X_r)$], leurs fréquences respectives absolues dans les différents vecteurs à n composantes (0, 1 ou E) représentant les différents cylindres ponctuels E_j, $1 \leqslant j \leqslant k$. Cette table est de dimension 2^r. La deuxième table est de dimension k ; elle donne, dans l'ordre, sous forme d'entiers, la suite des adresses du premier des 2^r vecteurs, puis celles du second ; et, ainsi de suite. De sorte que la complexité spatiale devient $2^r + k$.

IV. EVALUATION DE Ñ ET RECONNAISSANCE DE LA SATISFIABILITÉ DANS LE CAS D'UN MODÈLE ALÉATOIRE.

IV.1. Introduction ; description des différents modèles aléatoires.

Nous nous proposons de reprendre de façon conséquente, avec un point de vue que nous pensons plus global, l'aspect statistique considéré de façon originale dans [Simon & Dubois 1988], concernant l'évaluation approximative du nombre de solutions d'un problème SAT, dans le cadre d'un modèle aléatoire. Une telle évaluation correspondra à l'estimation par l'espérance mathématique. A cette fin, nous commencerons par préciser clairement les différents modèles aléatoires qu'on peut envisager et nous

nous appuyerons sur la formule d'inclusion et d'exclusion. D'autre part et de façon fondamentale, nous distinguerons dans ce cadre le problème probabiliste de la reconnaissance de la satisfiabilité ; ce qui permettra de clairement comprendre les résultats expérimentaux obtenus dans [Simon & Dubois 1988] sur le nombre moyen de clauses nécessaires pour atteindre l'insatisfiabilité.

Ce modèle aléatoire correspond en fait à celui de l'hypothèse d'absence de liaison que nous avons coutume de considérer en classification pour les comparaisons mutuelles entre parties d'un même ensemble Ω. Ici Ω est l'espace $\{0,1\}^n$ et les parties de Ω sont des cylindres ponctuels ; de sorte que nous serons conduits à introduire un modèle spécifique que nous avons déjà mentionné au paragraphe III.1.3., dont nous allons reprendre les éléments de façon plus systématique.

Relativement au couple (Ω, E) où $\Omega = \{0,1\}^n$ et E, un cylindre ponctuel d'ordre r, on peut oublier la structure cartésienne et seulement retenir que E est un sous ensemble de cardinal $\ell = 2^{n-r}$ d'un ensemble Ω de cardinal $m = 2^n$. Nous avons dans ces conditions dégagé trois formes fondamentales d'un modèle aléatoire associant au couple observé (Ω, E) un couple aléatoire (Ω^*, E^*) qui, chacune, d'une façon, respecte les caractéristiques cardinales de (Ω, E). Désignons par $({}_i\Omega^*, {}_iE^*)$ le couple aléatoire associé à la forme i, $1 \leqslant i \leqslant 3$.

Pour $i = 1$, ${}_i\Omega^* = \Omega$ et ${}_iE^*$ est une partie aléatoire de Ω de cardinal ℓ, élément de l'ensemble, muni d'une probabilité uniforme, des parties de Ω de même cardinal ℓ. Ce modèle respecte strictement

les caractéristiques cardinales $(2^n, 2^{n-r})$ de (Ω, E). En d'autres termes, en considérant le simplexe 2^Ω des parties de Ω, le modèle affecte toute la probabilité, en la répartissant uniformément, sur un même niveau ; celui défini par l'ensemble des parties de cardinal ℓ. Dans ces conditions, si E_0 est une partie fixée quelconque de Ω, de cardinal ℓ, on a :

$$Pr\left({}_1E^* = E_0\right) = \begin{cases} 0 & \text{si } card(E_0) \neq \ell \ ; \\ 1/\binom{m}{\ell} & \text{si } card(E_0) = \ell \ . \end{cases} \quad (1)$$

Pour $i = 2$, la mesure de probabilité, au lieu d'être affectée sur un seul niveau de l'ensemble 2^Ω des parties de Ω, elle sera ici répartie de façon plus diffuse sur la suite des niveaux du simplexe. Ainsi, la forme 2 du modèle aléatoire comporte deux pas ; le premier consiste dans le choix d'un niveau et le second, dans le choix d'un élément de ce niveau.

Pour le choix du niveau, considérons la variable aléatoire entière L, indice d'un niveau du simplexe 2^Ω et cardinal commun de toutes les parties représentées à ce niveau. On pose pour $Pr(L = \ell)$ la probabilité binomiale suivante :

$$Pr(L = e) = \binom{m}{e} \lambda^e (1-\lambda)^{m-e} \quad , (2)$$

où λ est la proportion $\ell/m = 2^{-r}$ (dans notre cas).

Pour le choix aléatoire d'un élément d'un même niveau d'indice e, la probabilité (2) affectée à ce niveau, est uniformément répartie sur l'ensemble des $\binom{m}{e}$ points (dont chacun représente une partie de cardinale);

chaque point sera de la sorte chargé de la probabilité $\lambda^e (1-\lambda)^{m-e}$. Ainsi, E_0 étant une partie fixée de Ω de cardinal c, on a :

$$Pr\{{}_2E^* = E_0 / card(E_0) = c\} = \lambda^c (1-\lambda)^{m-c} \quad (3)$$

La troisième forme du modèle aléatoire ($i=3$) comporte trois pas. Contrairement aux deux modèles précédents où Ω est fixé, on regardera ici Ω comme la réalisation d'un ensemble aléatoire Ω^* dont le cardinal $\mathcal{M}$ est une variable aléatoire de Poisson de paramètre m :

$$Pr\{\mathcal{M} = p\} = \frac{m^p}{p!} e^{-m} \quad , \quad (4)$$

pour tout p de l'ensemble des entiers.

Conditionnellement à une valeur p_0 de $\mathcal{M}$, on introduit un ensemble Ω_0 de cardinal p_0, qu'on peut d'ailleurs noter :

$$\Omega_0 = \{1, 2, \ldots, i, \ldots, p_0\} . \quad (5)$$

Les deux autres pas du modèle sont alors en tout point analogues à celui ($i=2$) précédent. Plus précisément, un sommet du niveau e du simplexe 2^{Ω_0} - représentant une partie de cardinal e de Ω_0 - $0 \leqslant e \leqslant p_0$, sera chargé de la probabilité $\lambda^e (1-\lambda)^{p_0-e}$, où le paramètre $\lambda = \ell/m$, conserve exactement le même sens que ci-dessus. Ainsi, le choix du niveau e du simplexe 2^{Ω_0} se fera avec la probabilité binomiale : $\binom{p_0}{e} \lambda^e (1-\lambda)^{p_0-e}$, $0 \leqslant e \leqslant p_0$.

Maintenant, imaginons que dans la construction du modèle aléatoire,

on veuille tenir compte de la structure cartésienne de (Ω, E) où Ω est l'espace $\{0,1\}^n$ et où E est un cylindre ponctuel d'ordre Les trois types de modèle aléatoires correspondants à $i = 1, 2$ et 3, peuvent très exactement être repris. Cependant au lieu de l'argument Ω, il s'agira de l'ensemble des composantes $\{1, 2, \ldots, i, \ldots, n\}$ et au lieu de E, il s'agira du sous ensemble $\{i_1, i_2, \ldots, i_r\}$ des r composantes instanciées pour E. D'autre part, si $\{1, 2, \ldots, j, \ldots, q\}$ est la réalisation de l'ensemble aléatoire associé à $\{1, 2, \ldots, i, \ldots, n\}$ et si $\{j_1, j_2, \ldots, j_t\}$ est la réalisation du sous ensemble de $\{1, 2, \ldots, j, \ldots, q\}$ associé à $\{i_1, i_2, \ldots, i_r\}$, alors, l'ensemble des 2^t cylindres ponctuels correspondants à toutes les instanciations possibles de la suite des variables $(X_{j_1}, X_{j_2}, \ldots, X_{j_t})$ est muni d'une probabilité uniforme; c'est à dire chacun des cylindres a la probabilité 2^{-t}.

Pour plus de précision et parce qu'il sera exploité ci-dessous après avoir été mentionné au paragraphe III. 1.3., considérons dans ce dernier contexte le troisième type de modèle aléatoire.

A n, on associe une variable aléatoire entière N de Poisson de paramètre n. Pour $N = q$, le deuxième pas du modèle consiste à choisir une partie aléatoire de l'ensemble $\{1, 2, \ldots, j, \ldots, q\}$ que nous notons ici $\{q\}$. La probabilité de tomber sur une partie de cardinal t ; c'est à dire, sur le niveau t du simplexe $2^{\{q\}}$ défini par l'ensemble des parties de $\{q\}$, est donnée par la probabilité binomiale:

$$\binom{q}{t} p^t (1-p)^{q-t} \quad , \quad (6)$$

où p est la fraction r/n.

Cette probabilité étant uniformément répartie sur le niveau t, la probabilité de tomber sur une partie spécifique $\{j_1, j_2, \ldots, j_t\}$ est $p^t(1-p)^{q-t}$.

Etant donnée la suite des composantes $\{j_1, j_2, \ldots, j_t\}$, il y a 2^t cylindres ponctuels d'ordre t, basés dans le sous espace des composantes $\{j_1, j_2, \ldots, j_t\}$ et tous sont équiprobables dans le modèle aléatoire.

Nous utiliserons selon les cas l'un ou l'autre des différents modèles aléatoires. Néanmoins, il est très intéressant de constater que l'influence du choix d'un modèle par rapport à un autre, est pratiquement indifférente relativement aux deux questions posées : estimation par la moyenne du nombre de solutions d'un problème SAT dans le cadre d'un modèle aléatoire de l'hypothèse d'absence de liaison et nombre moyen de clauses nécessaires dans le cadre d'un tel modèle pour atteindre l'insatisfiabilité.

IV.2 Nombre moyen de solutions d'un d'un problème SAT dans le cadre d'une hypothèse d'absence de liaison.

IV.2.1 Introduction.

Compte tenu du formalisme introduit dans le paragraphe II.4., relativement à l'ensemble des cylindres ponctuels $\{E_j / 1 \leq j \leq k\}$ de l'espace $\{0,1\}^n$, il y a lieu de donner une estimation de

$$\tilde{N} = \text{card}\left(\bigcup_{1 \leq j \leq k} E_j\right) \quad , (7)$$

(où $N = 2^n - \tilde{N}$) dans le cadre d'un modèle aléatoire traduisant l'indépendance probabiliste mutuelle entre les différents cylindres ponctuels et respectant de façon plus ou moins floue les caractéristiques cardinales des différents E_j, $1 \leqslant j \leqslant k$. Ce respect peut se faire dans le cadre d'un modèle aléatoire aveugle ou tenant compte de la structure cartésienne propre à un cylindre ponctuel.

Dans ces conditions, nous associerons à la suite $\{E_j / 1 \leqslant j \leqslant k\}$ - selon le même type de modèle aléatoire - une suite $\{E^*_j / 1 \leqslant j \leqslant k\}$ de k parties aléatoires indépendantes d'un même ensemble Ω^*, lui même associé à l'espace $\Omega = \{0,1\}^n$. Nous avons déjà vu que dans le cas où le modèle aléatoire n'est pas aveugle, une réalisation de Ω^* est un ensemble de la forme $\{0,1\}^q$ (avec $q = n$, pour deux des formes du modèle) et E^*_j, $1 \leqslant j \leqslant k$, est un cylindre ponctuel aléatoire de $\{0,1\}^q$. En fait nous avons pu nous rendre compte au paragraphe IV.1. ci-dessus qu'il y a six types de modèle aléatoire pour l'association

$$E_j \longrightarrow E^*_j \quad , \quad (8)$$

ce dernier ensemble aléatoire peut - pour préciser le type de modèle (cf. § IV.1.) - être noté ${}_1E^*_j$, ${}_2E^*_j$, ${}_3E^*_j$, ${}_1E'^*_j$, ${}_2E'^*_j$ et ${}_3E'^*_j$, $1 \leqslant j \leqslant k$, où ${}_iE'^*_j$ ($i = 1, 2$ ou 3) est un cylindre ponctuel aléatoire.

Quoiqu'il en soit et comme nous l'avons annoncé, le type de modèle n'influencera pas l'évaluation de :

$$\mathcal{E}(\tilde{N}^*) = \mathcal{E}\left[\operatorname{card}\left(\bigcup_{1 \leqslant j \leqslant k} E^*_j\right)\right] \quad , \quad (9)$$

où $\mathcal{E}$ désigne l'espérance mathématique.

Pour une telle évaluation et compte tenu de la linéarité de l'espérance, nous nous appuyerons sur la formule (23) (cf. § II.3), où il y a lieu de remplacer les E_j par les E_j^*, $1 \leqslant j \leqslant k$. Dans ces conditions, il suffira par conséquent de pouvoir évaluer :

$$\mathcal{C}\left[\operatorname{card}\left(E^*_{j_1} \cap E^*_{j_2} \cap \cdots \cap E^*_{j_h}\right)\right] \quad , \quad (10)$$

où $\{j_1, j_2, \ldots, j_h\}$ est une partie fixée de cardinal h, $1 \leqslant h \leqslant k$, de l'ensemble $\{1, 2, \ldots, k\}$ des indices.

En fait, nous irons jusqu'à donner pour l'un des modèles aléatoires, la loi de probabilité de la variable aléatoire :

$$Y(j_1, j_2, \ldots, j_h) = \operatorname{card}\left(E^*_{j_1} \cap E^*_{j_2} \cap \cdots \cap E^*_{j_h}\right) \quad , \quad (11)$$

(cf. § IV.2.4.).

Au paragraphe IV.2.2. nous rappellerons les lois de probabilité de $\operatorname{card}(G^* \cap H^*)$ où (G^*, H^*) est un couple de parties aléatoires indépendantes, associé à un couple observé de parties de Ω, conformément à l'un des modèles ($i = 1, 2$ ou 3) ci-dessus exprimés (cf. § IV.2.1). Nous préciserons d'autre part la loi de $\operatorname{card}(G^* \cap H^*)$ dans le cas où (G^*, H^*) est un couple de cylindres ponctuels aléatoires indépendants, associé à un couple de cylindres observé, dans le cadre du dernier modèle aléatoire mentionné au paragraphe IV.2.1.

Comme nous l'avons annoncé (cf. § IV.2.1) $\mathcal{C}[\operatorname{card}(G^* \cap H^*)]$ est invariant quelle que soit la forme du modèle aléatoire. Au paragraphe IV.2.3. nous déterminerons par récurrence l'espérance mathématique (10), qui possède également cette même propriété.

Nous pourrons alors déterminer $\mathcal{E}(\tilde{N})$ [cf. (9)] et nous retrouverons ainsi, de façon plus globale et plus directe, un résultat exprimé dans [Simon & Dubois 1988]. C'est de façon encore plus directe et plus ambitieuse que le résultat est obtenu au paragraphe IV.2.4.. Le paragraphe IV.2.5. signale une approche de type "processus de Markov" que nous a communiquée F. Daudé (chercheur en stage de thèse) et qui, également conduit au résultat.

IV.2.2. Loi de probabilité de card $(G^* \cap H^*)$.

(G, H) est un couple de parties de cardinaux respectifs g et h [card $(G) = g$ et card $(H) = h$] d'un ensemble Ω de cardinal m [card $(\Omega) = m$]. (G_1^*, H_1^*) est un couple de parties aléatoires indépendantes respectivement associées à (G, H) dans la forme $(i = 1)$ du modèle aléatoire (cf. § IV.1) : G_1^* (resp. H_1^*) est associé à G (resp. H). A l'indice brut :

$$s(G,H) = \text{card}\,(G \cap H) \quad , \quad (12)$$

déjà introduit ici dans un autre contexte [cf. (22) § III.1.3.], on associe l'indice brut aléatoire :

$$S_1(G,H) = \text{card}\,(G_1^* \cap H_1^*) \quad , \quad (13)$$

[cf. (23) § III.1.3.]. Nous établissons aisément [Lerman 1981] que $S_1(G,H)$ suit une loi de probabilité hypergéométrique de paramètres (m, g, h). Plus précisément :

$$Pr[S_1(G,H)=s] = \frac{\binom{g}{s}\binom{m-g}{h-s}}{\binom{m}{h}} = \frac{\binom{h}{s}\binom{m-h}{g-s}}{\binom{m}{g}} . \quad (14)$$

On établit que la moyenne et la variance de $S_1(G,H)$ sont respectivement, avec $\gamma = \frac{g}{m}$ (resp. $\eta = \frac{h}{m}$) et $\bar{\gamma} = 1-\gamma$ (resp. $\bar{\eta} = 1-\eta$) :

$$\mathcal{E}[S_1(G,H)] = m\gamma\eta \quad (15)$$

et

$$\text{var}[S_1(G,H)] = \frac{m^2}{(m-1)}\gamma\bar{\gamma}\eta\bar{\eta} . \quad (16)$$

Désignons maintenant par (G_2^*, H_2^*) le couple de parties aléatoires indépendantes associé à (G,H) dans le cadre de la forme $(i=2)$ du modèle aléatoire [cf. § IV.1]. Dans ce cas, la loi de

$$S_2(G,H) = \text{card}(G_2^* \cap H_2^*) , \quad (17)$$

est binomiale de paramètres $(m, \gamma\eta)$. Plus précisément, on a :

$$Pr[S_2(G,H)=s] = \binom{m}{s}\pi^s\bar{\pi}^{m-s} , \quad (18)$$

où nous avons noté $\pi = \gamma\eta$. On a alors :

$$\mathcal{E}[S_2(G,H)] = m\pi = m\gamma\eta \quad (19)$$

et

$$\text{var}[S_2(G,H)] = m\pi(1-\pi) = m\gamma\eta(1-\gamma\eta) . \quad (20)$$

Considérons à présent le triplet (Ω^*, G_3^*, H_3^*) dont les trois arguments sont aléatoires qu'on associe à (Ω, G, H) dans le cadre de la forme $(i=3)$ du modèle aléatoire [cf. § IV.1] et où G_3^* et

H_3^* sont indépendants. Dans ces conditions [Lerman 1981], la loi de probabilité de

$$S_3(G,H) = \text{card}\,(G_3^* \cap H_3^*) \quad , \quad (21)$$

est une loi de Poisson de paramètre $m\pi = m\gamma\eta$. Par conséquent :

$$\mathcal{E}[S_3(G,H)] = \text{var}[S_3(G,H)] = m\gamma\eta \quad . \quad (22)$$

Nous pouvons maintenant considérer les trois formes de l'hypothèse d'absence de liaison dans le cas où on tient compte de la structure cartésienne particulière de $\Omega = \{0,1\}^n$ et des cylindres ponctuels G et H [cf. fin du paragraphe IV.1.]. Nous pouvons désigner par $(G_i'^*, H_i'^*)$ le couple de cylindres ponctuels aléatoires indépendants, associé à (G, H) pour la forme i du modèle aléatoire, $i = 1, 2$ ou 3, qui, ici et encore une fois a pour ensemble fondamental :

$$\{m\} = \{1, 2, \ldots, \ell, \ldots, n\} \qquad (23)$$

et concerne le choix de deux parties indépendantes de l'ensemble correspondant à $\{m\}$, qui sont associées à $\{i_1, \ldots, i_r\}$ et $\{j_1, \ldots, j_s\}$, si le cylindre G (resp. H) est basé dans le sous espace étiqueté par $\{i_1, i_2, \ldots, i_r\}$ (resp. $\{j_1, j_2, \ldots, j_s\}$) ; ce qui signifie, l'instanciation des variables $\{X_{i_1}, X_{i_2}, \ldots, X_{i_r}\}$ (resp. $\{X_{j_1}, X_{j_2}, \ldots, X_{j_s}\}$) dans la clause que représente G (resp. H). Nous pourrions déterminer la loi de $S_i'(G,H)$

$$S_i'(G,H) = \text{card}\,(G_i'^* \cap H_i'^*) \quad , \quad (24)$$

pour chacun des trois modèles aléatoires ($i = 1, 2$ ou 3). Cependant, pour

des raisons de simplicité, parce que cela n'apportera pas davantage d'examiner les deux autres modèles et parce que le troisième modèle est particulièrement adapté (r et s "petits" devant n), nous nous contenterons d'étudier la loi de probabilité de

$$S'_3(G,H) = \text{card}\,(G'^{*}_3 \cap H'^{*}_3) \quad . \quad (25)$$

D'après ce qui précède, on peut noter (I^*_3, J^*_3) le couple de parties aléatoires indépendantes, d'une section commençante $\{q\} = \{1,2,\ldots,q\}$ de l'ensemble $\mathbb{N}$ des entiers, associé au couple :

$$[(i_1,\ldots,i_r),(j_1,\ldots,j_s)] \quad , \quad (26)$$

où q est la réalisation d'une variable aléatoire de Poisson $\mathcal{N}$ de paramètre n.

Toujours d'après ce qui précède, en désignant par ρ (resp. σ) le rapport r/n (resp. s/n), la variable aléatoire $\text{card}\,(I^*_3 \cap J^*_3)$ suit une loi de Poisson de paramètre $\mu = n\rho\sigma$. Très précisément, on a :

$$\Pr\{\text{card}\,(I^*_3 \cap J^*_3) = p\} = \frac{\mu^p}{p!}\, e^{-\mu} \quad , \quad (27)$$

p entier, positif ou nul.

Alors que, pour $\mathcal{N} = q$, $\text{card}\,(I^*_3)$ [resp. $\text{card}\,(J^*_3)$] suit une loi binomiale de paramètres (q,ρ) [resp. (q,σ)]. On en déduit que $\text{card}\,(I^*_3)$ [resp. $\text{card}\,(J^*_3)$] suit une loi de Poisson de paramètre r (resp. s).

Bien que nous ayons pu envisager les calculs dans toute leur complexité, nous allons simplifier en considérant le modèle où la loi

(27) de card $(I_3^* \cap J_3^*)$ est conservée ; mais où, avec la probabilité 1, $\mathcal{N} = n$, card $(I_3^*) = r$ et card $(J_3^*) = s$. Ce modèle d'approximation est d'autant plus excellent que – dans le contexte – r et s sont "petits" devant n.

Relativement à ce dernier modèle, désignons par K, la variable aléatoire $S'_3(G, H)$ [cf. (25) ci-dessus] et par P celle, card $(I_3^* \cap J_3^*)$. Une valeur 0 de P, conduit nécessairement à la valeur suivante de K ;

$$2^{n-r-s}$$

qui est seulement atteinte dans ce cas. On a donc :

$$Pr\{K = 2^{n-r-s}\} = e^{-\mu} \quad . \qquad (28)$$

Une valeur p de P peut conduire soit à une valeur 0 de K, en cas de disjonction entre les deux cylindres ponctuels aléatoires G'^*_3 et H'^*_3, soit à une valeur $2^{n-r-s+p}$ en cas d'inclusion de l'un des cylindres ponctuels dans l'autre. Pour p fixé, la première circonstance a lieu avec la probabilité $(1 - 2^{-p})$; et la seconde, avec la probabilité complémentaire 2^{-p}. Finalement et en tenant compte du caractère d'approximation du modèle :

$$Pr\{K = 0\} = \sum_{h \neq 0} \frac{\mu^p}{p!} e^{-\mu} \cdot (1 - 2^{-p})$$

$$= 1 - e^{-\mu/2} \qquad (29)$$

et

$$Pr\{K = 2^{n-(r+s-p)}\} = \frac{\mu^p}{p!} e^{-\mu} \cdot 2^{-p} = e^{-\mu/2} \cdot \left[\frac{(\mu/2)^p}{p!} e^{-\mu/2}\right], \quad (30)$$

où $p \geq 1$.

Les calculs de l'espérance mathématique et du moment absolu d'ordre 2 de la variable aléatoire K, donnent très exactement :

$$\mathcal{E}(K) = 2^{n-r-s} \qquad (31)$$

et

$$\mathcal{E}(K^2) = 2^{2(n-r-s)} \cdot e^{\mu} \; ; \; (32)$$

de sorte que

$$var(K) = 2^{2(n-r-s)} \left(e^{\mu} - 1\right) \; . \; (33)$$

Ainsi se trouve justifié l'expression - conçue dans un tout autre contexte - (33) de $R(g,h)$ [cf. § III.1.3.].

D'autre part et comme nous l'avons bien annoncé, la moyenne $\mathcal{E}(K)$ ci-dessus est invariante quelle que soit la forme du modèle aléatoire considérée. En effet, en considérant la valeur commune $m\gamma\eta$ de (15), (19) ou (22) ci-dessus, on a bien :

$$m\gamma\eta = 2^{n-r-s} \; , \; (34)$$

si $m = 2^n$, $\gamma = 2^{n-r}/2^n = 2^{-r}$ et $\eta = 2^{n-s}/2^n = 2^{-s}$.

IV.2.3. Evaluation par récurrence de $\mathcal{E}[\,card(E_1^* \cap E_2^* \cap \dots \cap E_h^*)\,]$ et calcul de $\mathcal{E}(\tilde{N}^*)$.

Rappelons que $(E_1^*, E_2^*, \dots, E_h^*)$ est une suite de cylindres ponctuels aléatoires indépendants associée à la suite observée $(E_1, E_2, \dots, E_h)$ dans le cadre d'un même modèle aléatoire [cf. § IV.1. et IV.2.1.].

On a ainsi à évaluer une expression telle que (10) ci-dessus (§ IV.2.1.). Supposons que E_j, $1 \leqslant j \leqslant h$, est un cylindre ponctuel d'ordre r_j (r_j variables sont instanciées) ; dont le volume ou cardinal est par conséquent 2^{n-r_j}. On suppose de plus avoir établi que pour $2 \leqslant i \leqslant j-1$,

$$\mathcal{E}\left[\operatorname{card}\left(E_1^* \cap E_2^* \cap \cdots \cap E_i^*\right)\right] = 2^{n-(r_1+r_2+\cdots+r_i)} \qquad (35)$$

et nous allons démontrer que

$$\mathcal{E}\left[\operatorname{card}\left(E_1^* \cap E_2^* \cap \cdots \cap E_j^*\right)\right] = 2^{n-(r_1+r_2+\cdots+r_j)} \quad . \quad (36)$$

Or on a

$$\mathcal{E}\left[\operatorname{card}\left(E_1^* \cap E_2^* \cap \cdots \cap E_j^*\right)\right] =$$

$$\sum_s \mathcal{E}\left[\operatorname{card}\left(E_1^* \cap \cdots \cap E_{j-1}^* \cap E_j^*\right) \Big/ \operatorname{card}\left(E_1^* \cap \cdots \cap E_{j-1}^*\right) = 2^{n-s}\right]$$

$$\times \Pr\left[\operatorname{card}\left(E_1^* \cap \cdots \cap E_{j-1}^*\right) = 2^{n-s}\right] . \quad (37)$$

En effet, $E_1^* \cap \cdots \cap E_{j-1}^*$ est un cylindre ponctuel, dont le volume est nécessairement de la forme 2^{n-s}.

Compte tenu de ce qui a été établi au paragraphe IV.2.2., la valeur de l'espérance mathématique qui apparaît dans l'expression (37) est :

$$2^{n-s-r_j}$$

Ainsi, l'expression (37) se réduit à

$$2^{-r_j}\, \mathcal{E}\left[\operatorname{card}\left(E_1^* \cap \cdots \cap E_{j-1}^*\right)\right] . \quad (38)$$

Compte tenu de l'hypothèse de récurrence [cf. (35) ci-dessus] qui a bien été établie jusqu'à 2 [cf. § IV.2.2. ci-dessus], on a le résultat général.

Ce résultat nous permet d'evaluer $\mathcal{C}(\tilde{N}^*)$ [cf. (9) § IV.2.1. ci-dessus]. Compte tenu de la formule d'inclusion et d'exclusion, ainsi bien sûr que de la linéarité de l'espérance :

$$\mathcal{C}(\tilde{N}^*) = \sum_{1 \leqslant h \leqslant k} (-1)^{h+1} \sum_{\{j_1, \ldots, j_h\}} 2^{n-(r_{j_1}+\cdots+r_{j_h})} \quad , \qquad (39)$$

où r_{j_i}, $1 \leqslant i \leqslant h$, est l'ordre du cylindre E_{j_i}.

Le développement du second membre de (39) nous permet de reconnaître que

$$\mathcal{C}(\tilde{N}^*) = 2^n \left[1 - \prod_{1 \leqslant j \leqslant k} (1 - 2^{-r_j}) \right] \quad , (40)$$

qui correspond exactement au résultat établi dans [Simon & Dubois 1988] (qui notent $\tilde{N}$ pour N^*; alors que nous avons réservé le symbole $\sim$ pour la complémentation).

Les auteurs que nous venons de référencer s'étonnent de ce que

$$\mathcal{C}(N^*) = \left[\prod_{1 \leqslant j \leqslant k} (1 - 2^{-r_j}) \right] 2^n \quad , (41)$$

ne dépende pas de la distribution des différentes variables X_i, $1 \leqslant i \leqslant n$, dans la suite des clauses. En vérité, ce résultat est conceptuellement, parfaitement prévisible ; puisque le modèle aléatoire de l'hypo-

thèse d'absence de liaison ne fait intervenir que le nombre de variables instanciées par clause et non le nombre de clauses où une même variable se trouve instanciée. Un modèle aléatoire où il s'agit de tenir à la fois compte du nombre de variables instanciées par clause et du nombre de clauses où une même variable se trouve instanciée, peut avoir comme argument aléatoire un tableau d'incidence (formé de zéros et de uns) à k lignes et n colonnes, dont les marges sont fixées. Une marge ligne (resp. colonne) représente le nombre de composantes égales à 1 dans la ligne (resp. colonne). Une valeur égale à 1 à la j-ème ligne et i-ème colonne ($1 \leqslant j \leqslant k$, $1 \leqslant i \leqslant n$) indiquera que la variable X_i doit être instanciée dans la j-ème clause z alors qu'une valeur égale à 0, indiquera le contraire.

L'étude beaucoup plus complexe d'un tel modèle aléatoire aurait pu conduire à une valeur de $\mathcal{O}(N^*)$ qui dépend des marges du tableau d'incidence aléatoire, donc aussi d'un aspect de la distribution des variables X_i, $1 \leqslant i \leqslant n$. Cependant, on peut se demander — par rapport au problème qui nous préoccupe — si les résultats escomptés seront à la hauteur de l'effort nécessaire.

A partir de maintenant et sans restreindre la généralité par rapport au problème de résolution de la complexité, nous allons supposer que tous les r_j, $1 \leqslant j \leqslant k$, sont égaux entre eux à r où $r \geqslant 3$, puisqu'on démontre que pour $r < 2$, la complexité du problème SAT est polynomiale. Dans ces conditions, la relation (41) ci-dessus devient :

$$\mathcal{C}(N^*) = (1-2^{-r})^k \times 2^n \quad , \quad (42)$$

qui est une fonction décroissante par rapport à k et croissante par rapport à r. Inversement, relativement au chargement moyen de l'espace complémentaire des solutions :

$$\mathcal{C}(\tilde{N}^*) = [1-(1-2^{-r})^k] \times 2^n \quad , \quad (43)$$

il s'agit d'une fonction croissante par rapport à k et décroissante par rapport à r.

Revenons maintenant sur l'amplitude (17) du paragraphe III.1.2, de l'encadrement (29) du paragraphe II.3. et calculons son espérance mathématique dans l'hypothèse d'absence de liaison correspondante à l'un des modèles aléatoire ci-dessus introduits (cf. § IV.1). On a, compte tenu de l'expression (36) ci-dessus,

$$\mathcal{C}(S^*_{2p-1} - S^*_{2p}) = \binom{k}{2p} 2^{n-2pr} \quad . \quad (44)$$

En indiquant par M_{2p} l'expression (44) ci-dessus, on obtient

$$\frac{M_{2(p+1)}}{M_{2p}} = 2^{-2r} \times \frac{(k-2p)(k-2p-1)}{(2p+1)(2p+2)} \quad . \quad (45)$$

On peut établir que le second membre est strictement inférieur à l'unité si p est supérieur ou égal à $k/2^{r+1}$. Ainsi, plus r est grand, davantage l'espérance de l'amplitude de l'encadrement diminue, dès que p dépasse une fraction d'autant plus petite de k. Une "grande" valeur de r suppose une "petite" valeur de $\mathcal{C}(\tilde{N}^*)$.

IV.2.4. Loi de probabilité de $\mathrm{card}(E_1^* \cap \cdots \cap E_h^*)$.

Nous allons à présent déterminer dans le cadre d'une hypothèse d'absence de liaison qui utilise la forme ($i=1$) du modèle aléatoire de choix [cf. § IV.1], directement, la loi de probabilité de $\mathrm{card}(E_1^* \cap \cdots \cap E_h^*)$. Précisons qu'ici $(E_1^*, \ldots, E_h^*)$ est une suite de parties aléatoires indépendantes de même cardinal ℓ qui peut correspondre au volume 2^{n-r} d'un même cylindre ponctuel. Ces dernières sont des éléments de l'ensemble $\mathcal{P}_\ell(\Omega)$ des parties de même cardinal ℓ de l'ensemble total Ω dont on notera par m le cardinal; Ω peut correspondre à l'espace $\{0,1\}^n$ où alors $m = 2^n$. $\mathcal{P}_\ell(\Omega)$ comporte $\binom{m}{\ell}$ éléments. On a :

Propriété 1 : Pour l'entier p compris dans l'intervalle

$$[\max(0, h\ell - m)\ ,\ \ell\ ],$$

$$\Pr\{\mathrm{card}(E_1^* \cap \cdots \cap E_h^*) \geqslant p\} \cong \binom{\ell}{p}\left[\binom{\ell}{p} \Big/ \binom{m}{p}\right]^{h-1}. \quad (46)$$

Le résultat donné au second membre de (46) correspond à une approximation. Nous nous en contenterons car un résultat exact nous paraît inextricable à obtenir. Conditionnellement au choix E_1° de la partie aléatoire E_1^*, la condition nécessaire et suffisante pour que :

$$\mathrm{card}(E_1^* \cap \cdots \cap E_h^*) \geqslant p \quad , \quad (47)$$

est qu'il existe une partie de cardinal p de E_1° qui soit incluse

dans chacune des parties aléatoires $E_2^*, E_3^*, \ldots, E_h^*$. Si une telle partie est spécifiée, la probabilité pour qu'il en soit ainsi est :

$$\left[\frac{\binom{\ell}{n}}{\binom{m}{n}}\right]^{h-1} \quad . \quad (48)$$

Toutefois, cette partie n'a pas à être spécifiée et il y en a $\binom{\ell}{n}$ et, en toute rigueur, on aurait à appliquer la formule d'inclusion et d'exclusion, car plusieurs parties de cardinal p de E_1° peuvent être incluses dans un même ensemble aléatoire E_j^*, $1 \leqslant j \leqslant h$. En négligeant la probabilité d'au moins deux parties dans le cadre de la probabilité d'inclusion d'au moins une partie, on obtient le résultat annoncé ci-dessus [cf. (46)].

Ce résultat est une approximation par excès de la probabilité définie au premier membre de (46). Une approximation par défaut du second membre de (46), qui est d'autant plus valable que p est "petit", est donnée par :

$$\binom{\ell}{n} \lambda^n \quad , \quad (49)$$

où nous notons λ la quantité $(\ell/m)^{h-1}$.

Dans ces conditions, Y désignant la variable aléatoire définie par $\operatorname{card}(E_1^* \cap \ldots \cap E_h^*)$, on admettra d'autant mieux

$$Pr\{Y \geqslant n\} = \binom{\ell}{n} \lambda^n \quad , \quad (50)$$

qu'on définit ainsi une variable aléatoire (la somme des probabilités

est égale à 1) et que de plus, on retrouve dans ce dernier cadre que $\mathcal{E}(Y)$ est bien égal à la valeur correspondante à (36) ci-dessus ; c'est-à-dire, avec les notations que nous avons adoptées $m\lambda$. On suppose la restriction $h-1 \geqslant (\mathrm{Log}_2 \ell)/(\mathrm{Log}_2 (m/\ell))$.

Propriété 2 : La suite des nombres positifs $\{[\binom{\ell}{n}\lambda^n - \binom{\ell}{n+1}\lambda^{n+1}] / 0 \leqslant n \leqslant \ell\}$ définit la loi de probabilité d'une variable aléatoire Y de moyenne $\mathcal{E}(Y)$ égale à $m\lambda$.

Si $m = 2^n$ et $\ell = 2^{n-r}$, la restriction mentionnée ci-dessus correspond à l'inégalité :

$$n \leqslant hr \quad . \quad (51)$$

Cette inégalité assure que tous les termes de la suite dans la propriété ci-dessus, sont positifs. La somme de ces termes prend la forme :

$$\sum \{Pr\{Y = n\} / 0 \leqslant n \leqslant \ell\} = (1+\lambda)^{\ell} - [(1+\lambda)^{\ell} - 1]$$

$$= 1. \quad (52)$$

D'autre part, on a :

$$\mathcal{E}(Y) = \sum_{n \geqslant 1} Pr\{Y \geqslant n\} \quad . \quad (53)$$

En tenant compte de (50) ci-dessus, on obtient :

$$\mathcal{E}(Y) = (1+\lambda)^{\ell} - 1$$

$$\simeq 1 + \ell \times \left(\frac{\ell}{m}\right)^{h-1} - 1 = \ell \times \left(\frac{\ell}{m}\right)^{h-1}, \quad (54)$$

ce qui se traduit par 2^{n-hr}, si $m = 2^n$ et $\ell = 2^{n-r}$. C.Q.F.D.

Remarques. Les résultats que nous avons obtenus dans le cadre des propriétés 1 et 2 ci-dessus et qui correspondent à des approximations sont surtout valables dès que h est assez grand. On atteint ainsi la cohérence pour $h \geq n/r$, en ce qui concerne la propriété 2. Signalons que dans un tout autre contexte (publication interne Irisa, 1984) nous avions déterminé très exactement la loi de probabilité de card $(E_1^* \cap E_2^* \cap E_3^*)$ (soit pour $h=3$). Si maintenant on considère un modèle aléatoire plus "flou", où au lieu de E_j^*, on considère une suite de ℓ points aléatoires indépendants de Ω, $1 \leq j \leq h$. Dans ces conditions, la probabilité pour qu'un point donné de Ω tombe dans l'intersection des h suites aléatoires indépendantes est $(\ell/m)^h$. On peut alors considérer, que le nombre P' de points de Ω tombant dans la précédente intersection, suit une loi de Poisson [m "grand" et $(\ell/m)^h$ "petit"] de paramètre $\ell\lambda$, où $\lambda = (\ell/m)^{h-1}$. L'espérance mathématique de P' est bien égale à 2^{m-rh}, pour $m = 2^n$ et $\ell = 2^{n-r}$.

IV.2.5. Une interprétation en termes de chaîne de Markov.

Une telle interprétation nous a été communiquée par F. Daudé (chercheur-thésard). En étant plus systématique, elle s'apparente d'une certaine façon dans son esprit à l'approche de [Simon & Dubois 1988]; en ce sens où on s'intéresse à l'accroissement aléatoire du chargement d'un ensemble Ω

de taille m, entre deux tirages consécutifs, d'une suite de tirages aléatoires indépendants, de parties de Ω de cardinal ℓ ($\ell < m$). Répétons une fois de plus, que dans notre cas, on peut considérer $m = 2^n$ et $\ell = 2^{n-r}$.

Pour faire image, on peut se figurer une urne qui ne comprend au départ que des boules noires. La contenance de cette urne reste invariablement égale à m au cours du processus. A la h-ème étape on extraie, uniformément au hasard, un sous ensemble de ℓ boules et on remet à leur place dans l'urne ℓ boules rouges. Ainsi, à la première étape ($h = 1$) ce sont ℓ boules noires qui se trouvent remplacées par ℓ boules rouges ; mais, à une étape courante, c'est un mélange de boules noires ou rouges, auquel on substitue des boules exclusivement rouges. Dans ces conditions, on s'intéresse au taux de remplissage par des boules rouges au bout de k tirages aléatoires indépendants.

En désignant par S_h la variable aléatoire entière définie par le nombre de boules rouges dans l'urne après la h-ème étape. On a pour $Pr(S_h = x+u \,/\, S_{h-1} = x)$, la probabilité hypergéométrique:

$$Pr(S_h = x+u \,/\, S_{h-1} = x) = \frac{\binom{m-x}{u}\binom{x}{\ell-u}}{\binom{m}{\ell}} \quad (55)$$

Ce qui conduit à

$$\mathcal{E}(S_h - S_{h-1} \,/\, S_{h-1}) = \frac{(m - S_{h-1})\ell}{m} \; ; \quad (56)$$

d'où :

$$\mathcal{E}(S_h / S_{h-1}) = S_{h-1} - \frac{\ell}{m} + \ell \qquad (57)$$

Comme $\mathcal{E}(S_h / S_{h-1})$ ne dépend que de S_{h-1}, on a :

$$\mathcal{E}(S_h) = \mathcal{E}[\mathcal{E}(S_h / S_{h-1})] . \qquad (58)$$

Il en résulte une récurrence qui conduit à

$$\mathcal{E}(S_h) = m\left[1 - \left(1 - \frac{\ell}{m}\right)^h\right] \qquad (59)$$

$$= 2^n \left[1 - (1 - 2^{-r})^h\right] , \qquad (60)$$

pour $m = 2^n$ et $\ell = 2^{n-r}$, ce qui correspond exactement à la formule (43) ci-dessus.

La démarche ci-dessus de F. Daudé nous permet de nous rendre compte de toute la complexité de la loi de probabilité de S_k. Nous obtenons en effet, en écrivant S_k sous la forme :

$$S_k = \ell + T , \qquad (61)$$

où ℓ correspond à la valeur de S_1 et où $k \geqslant 2$,

$$Pr\{S_k = \ell + t\} = \sum \left\{ \left[\frac{(m-\ell)!}{t_2!\, t_3! \ldots t_k!\, (m-\ell-t_2-t_3-\cdots-t_k)!} \right] \Big/ \frac{}{\binom{m}{\ell}^{k-1}} \right.$$

$$\left. (t_2, t_3, \ldots, t_k) \in \mathcal{P}_{k-1}(t) \right\} . \qquad (62)$$

où $\mathcal{P}_{k-1}(t)$ est l'ensemble de tous les $(k-1)$ uples ordonnés d'entiers dont chacun est inférieur ou égal à ℓ et dont la somme est égale à t. On peut remarquer que l'expression (62) peut aussi se mettre sous la forme :

$$Pr\{S_k = \ell + t\} = \sum_{(t_2, t_3, \ldots, t_k) \in \mathcal{P}_{k-1}(t)} \left\{ \binom{m-\ell}{t} \times \binom{t}{t_2\ t_3 \ldots\ t_k} \Big/ \binom{m}{\ell}^{k-1} \right\}, \quad (63)$$

où $\binom{t}{t_2\ t_3 \ldots\ t_k}$ désigne un coefficient multinomial. Nous pouvons espérer à l'avenir exploiter plus avant cette expression, notamment pour déterminer la variance de S_k, ou d'ailleurs de $(S_k - \ell)$, par le calcul du moment factoriel d'ordre 2 : $\mathcal{E}[(S_k - \ell)(S_k - \ell - 1)]$.

IV.3. Reconnaissance de la satisfiabilité ; nombre moyen de clauses pour atteindre l'insatisfiabilité dans le cadre d'un modèle aléatoire.

IV.3.1. Introduction

Nous allons reprendre le contexte ci-dessus ; c'est à dire, la forme 1 de l'hypothèse d'absence de liaison [cf. (1) § IV.1]. Relativement à notre expression du problème, commençons par reprendre (9) du paragraphe IV.2.1. dont l'évaluation a déjà été

donnée ci-dessus [cf. (43), (59) ou (60) ci-dessus], dans le cas que nous supposons où toutes les parties aléatoires E_j^*, $1 \leqslant j \leqslant k$, sont de même cardinal ℓ, qu'on pourra - pour tenir compte du contexte - noter 2^{n-r} ; alors que le cardinal m de l'ensemble Ω peut être écrite 2^n.

Précisons le premier membre de l'expression (9) en le notant $\mathcal{C}(\tilde{N}_k^*)$ et intéressons nous - à la manière de F. Daudé (communication personnelle) - à la valeur k_0 de k pour laquelle l'espérance mathématique du chargement, au bout de k sous ensembles aléatoires E_j^*, s'écarte de moins d'une unité du cardinal m de l'ensemble Ω ; soit :

$$\mathcal{C}(\tilde{N}_k^*) = m\left[1-\left(1-\frac{\ell}{m}\right)^k\right] > m-1 \quad . \quad (64)$$

On obtient

$$k > \frac{l_n\, m}{l_n\, m - l_n(m-\ell)} \quad , \quad (65)$$

ce qui correspond avec les notations ci-dessus à

$$\frac{k}{m} > \frac{-l_n 2}{l_n(1-2^{-r})} \simeq 2^r\, l_n 2 \quad ; \quad (66)$$

où la dernière approximation est d'autant plus précise que r est grand.

On retrouve ainsi exactement la valeur

$$k_0 = m \times \left[\frac{-l_n 2}{l_n(1-2^{-r})}\right] \simeq m \times 2^r\, l_n 2 \quad , \quad (67)$$

considérée dans [Simon & Dubois 1988], pour laquelle on peut croire qu'elle fournit la moyenne du nombre de clauses aléatoires permettant d'atteindre, dans le cadre du modèle aléatoire considéré, l'insatisfiabilité. Par rapport à notre expression du problème, il s'agirait du nombre moyen de parties aléatoires et indépendantes E_j^*, permettant le remplissage total de Ω.

En vérité il n'en est rien car la variable aléatoire qui nous intéresse ici est de nature essentiellement différente. En effet, relativement à la suite

$$\{ E_j^* / j \geq 1 \} \quad , \quad (68)$$

de parties aléatoires et indépendantes de même cardinal ℓ de Ω, nous définissons la variable aléatoire K dont une réalisation correspond à la plus petite valeur k de l'indice j, pour laquelle on a (par conséquent pour la première fois) :

$$\bigcup \{ E_j / 1 \leq j \leq k \} = \Omega \quad , \quad (69)$$

où E_j est la réalisation de E_j^*, $1 \leq j \leq k$.

Ce qui nous intéresse ici est alors l'espérance mathématique de K. Il s'agit par rapport à l'image du paragraphe IV.2.5. du nombre moyen de tirages nécessaires pour que l'urne ne contienne plus que des boules rouges ; et non, de la valeur de k pour laquelle le chargement moyen (par des boules rouges) s'écarte

de moins d'une unité du chargement total. On se rendra clairement compte que $\mathcal{C}(K)/n$ est sensiblement inférieur à k_0/n ; ce qui explique les résultats de la table 1 du paragraphe " Experimental verification " de [Simon & Dubois 1988].

Notre problème est en fait une généralisation de celui dit " des temps d'attente dans un échantillonnage avec remise " [cf. Feller 1964, chap. IX, § 3. d.] où $\ell = 1$, ce qui correspond à $r = n$. Dans ce cas, on obtient, conformément au second membre de (67) :

$$k_0(n, r=n) = 2^n \ln 2^n . \quad (70)$$

A rigoureusement parler, la même erreur d'interprétation subsiste dans la référence qu'on vient de mentionner. Toutefois, l'approximation de $\mathcal{C}(K)$ par k_0 est bonne dans ce cas car $\ell = 1$. Et, de façon générale, comme le laisse pressentir la table ci-dessus mentionnée, l'approximation par excès de $\mathcal{C}(K)$ par k_0, est d'autant meilleure que r/n est grand ; en rappelant que $\ell = 2^{n-r}$.

IV. 3. 2. *Moyenne du nombre de cylindres ponctuels aléatoires permettant la couverture de l'espace total.*

On reprend (68) ci-dessus où on suppose que E_j^* est un cylindre ponctuel aléatoire dont exactement r composantes

sont instanciées et donc de volume 2^{n-r}. Soit $\alpha=(\alpha_1,\dots,\alpha_i,\dots,\alpha_n)$ un point particulier de l'espace $\Omega=\{0,1\}^n$. Nous allons commencer par préciser quelle est la probabilité pour qu'un cylindre ponctuel E^* aléatoire dont $\{E^*_j / j\geq 1\}$ est une suite de copies indépendantes, n'atteigne pas le point α. Pour cela, il faut et il suffit que l'une des composantes aléatoires spécifiées dans E^* soit différente de celle, correspondante, de α. En utilisant la formule d'inclusion et d'exclusion, on a :

$$Pr\{\alpha \notin E^*\} = r\times\frac{1}{2} - \binom{r}{2}\times\left(\frac{1}{2}\right)^2 + \dots + (-1)^{r+1}\binom{r}{r}\times\left(\frac{1}{2}\right)^r$$

$$= 1-\frac{1}{2^r} = 1-\frac{2^{n-r}}{2^n} = 1-\frac{\ell}{m}. \quad (71)$$

$1-(\ell/m)$ est aussi la probabilité pour qu'une partie aléatoire E^* de cardinal ℓ d'un ensemble Ω de cardinal m, définie dans le cadre de la première forme du modèle aléatoire [cf. (1) § IV.1.], ne couvre pas un point particulier de Ω. De sorte qu'on ne restreint en rien la généralité en supposant que $\{E^*_j / j\geq 1\}$ est une suite de parties aléatoires indépendantes de même cardinal ℓ, définies dans le cadre de la première forme du modèle aléatoire. On se réfèrera à $m=2^n$ et $\ell=2^{n-r}$; de sorte que $m=2^r\ell$.

K désignant la variable aléatoire définie au paragraphe IV.3.1. ci-dessus et dont il s'agit de calculer l'espérance mathématique $\mathcal{E}(K)$, nous allons commencer par donner un développement qui exprime $Pr\{K>k\}$.

A cette fin, désignons par q_s^k, la probabilité pour qu'au bout de k parties aléatoires E_j^*, un sous ensemble spécifié de Ω comprenant s éléments, ne soit pas touché par $U\{E_j^* / 1 \leqslant j \leqslant k\}$. On a :

$$q_s^k = \left[\binom{m-s}{\ell} \Big/ \binom{m}{\ell} \right]^k = \left[\frac{(m-\ell)(m-\ell-1)\cdots(m-\ell-s+1)}{m\,(m-1)\cdots(m-s+1)} \right]^k . \quad (72)$$

q_s tend vers zéro lorsque s croît ; et ce, d'autant plus, que r est grand. D'autre part, une valeur approximative de q_s, d'autant plus précise que s est petit, est donnée par :

$$q_s = q^s, \text{ où } q = (m-\ell)/m = 1 - 2^{-r}. \quad (73)$$

En désignant par A_ω, l'évènement : le point ω de Ω n'est pas couvert par $U\{E_j^* / 1 \leqslant j \leqslant k\}$; l'évènement $\{K > k\}$ correspond à la réalisation de l'un ou l'autre des évènements A_ω. La formule d'inclusion et d'exclusion permet d'écrire :

$$\Pr\{K > k\} = m\, q_1^k - \binom{m}{2} q_2^k + \cdots + (-1)^{s+1} \binom{m}{s} q_s^k + \cdots$$
$$\cdots + (-1)^{m-\ell+1} \binom{m}{m-\ell} q_{m-\ell}^k . \quad (74)$$

Le terme général du développement ci-dessus peut se mettre sous la forme suivante :

$$\binom{m}{s} q_s^k = \binom{m-\ell}{s} q_s^{k-1} . \quad (75)$$

De la sorte, l'expression (74) ci-dessus peut s'écrire :

$$Pr\{K > k\} = \binom{m-\ell}{1} q_1^{k-1} - \binom{m-\ell}{2} q_2^{k-1} + \cdots + (-1)^{s+1} \binom{m-\ell}{s} q_s^{k-1}$$
$$+ \cdots + (-1)^{m-\ell+1} \binom{m-\ell}{m-\ell} q_{m-\ell}^{k-1} \quad . \quad (76)$$

On doit pouvoir établir que pour $k < m/\ell = 2^r$, la probabilité (76) ci-dessus est égale à 1. Nous l'avons en tout cas vérifié pour $m = 8$, $\ell = 2$ et $k = 2$.

En admettant l'approximation par excès (73) de q_s par q^s, on a :

$$Pr\{K \leqslant k\} \simeq \left(1 - q^{k-1}\right)^{m-\ell} \quad , \quad (77)$$

qui tend bien vers 1, pour k tendant vers l'infini.

En exploitant l'expression $\mathcal{E}(K) = \sum_k \{ Pr\{K \geqslant k\} / k \geqslant 1 \}$, on a à sommer l'expression (76) à partir de $k = 0$, pour obtenir :

$$\mathcal{E}(K) = \binom{m-\ell}{1} \frac{1}{q_1(1-q_1)} - \binom{m-\ell}{2} \frac{1}{q_2(1-q_2)} + \cdots$$
$$\cdots + (-1)^{s+1} \binom{m-\ell}{s} \frac{1}{q_s(1-q_s)} + \cdots$$
$$\cdots + (-1)^{m-\ell-1} \binom{m-\ell}{m-\ell} \frac{1}{q_{m-\ell}(1-q_{m-\ell})} \quad . \quad (78)$$

On a également :

$$\mathcal{E}(K) = \sum_{1 \leqslant s \leqslant m-\ell} (-1)^{s+1} \binom{m}{s} \frac{1}{1-q_s} \quad . \quad (79)$$

dont on peut considérer l'approximation à partir de (73) :

$$\mathcal{G}(K) = \sum_{1 \leqslant s \leqslant m-\ell} (-1)^{s+1} \binom{m}{s} \frac{1}{1-q^s} \quad . \quad (80)$$

Malgré leur complexité (développement de $m-\ell$ termes), les formules (79) et (80) méritent d'être expérimentées ; ce qui peut être fait jusqu'à un niveau non négligeable, compte tenu de la puissance actuelle des ordinateurs. A cette fin, on déterminera le coefficient binomial :

$$\binom{m}{s} = \prod_{0 \leqslant i \leqslant s-1} \left(\frac{m-i}{s-i} \right) \quad , \quad (81)$$

à partir de son logarithme.

IV 3.3. *Une formule de récurrence pour la loi de K.*

E_1^* couvre nécessairement un sous ensemble de cardinal ℓ de Ω. k étant supérieur ou égal à 3, désignons par $P_m(k-1, \underline{u})$ la probabilité de couverture exacte d'un sous ensemble <u>spécifié</u> de Ω, de cardinal u. On a :

$$\ell \leqslant u \leqslant \min [(k-1)\ell, m] \quad . \quad (82)$$

Si maintenant on désigne par $Q_m(k-1, u)$, la probabilité de couverture exacte d'un sous ensemble <u>non spécifié</u> de Ω, de cardinal u, $\underline{u}$. On a :

$$Q_m(k-1, u) = \binom{m}{u} P_m(k-1, \underline{u}) \quad , \quad (83)$$

où le premier facteur du second membre correspond au choix de la partie $\underline{u}$ de cardinal u. On a d'autre part :

$$P_m(k-1, \underline{u}) = \left[\frac{\binom{u}{\ell}}{\binom{m}{\ell}}\right]^{k-1} Q_u(k-1, u) \quad ; \quad (84)$$

il s'agit en effet que les $(k-1)$ parties tombent dans $\underline{u}$ et que, sachant cela, il y ait couverture totale de $\underline{u}$.

On a la formule de récurrence :

$$Q_m(k, m) = \sum \left\{ \frac{\binom{u}{\ell - m + u}}{\binom{m}{\ell}} \times Q_m(k-1, u) \Big/ m - \ell \leqslant u \leqslant m \right\}. \quad (85)$$

En effet, la partie E_k^* de cardinal ℓ doit être choisie de façon à recouvrir la partie laissée vide de Ω de cardinal $(m-u)$ et le choix concerne $\ell - (m-u)$ éléments parmi les u éléments de $\underline{u}$.

En reprenant $Q_m(k-1, u)$ à partir de (83) et de (84) et en effectuant une certaine mise en forme, on peut écrire :

$$Q_m(k, m) = \sum \left\{ \binom{\ell}{v} \times \left[\frac{\binom{m-v}{\ell}}{\binom{m}{\ell}}\right]^{k-1} Q_u(k-1, u) \Big/ 0 \leqslant v < \ell \right\}, \quad (86)$$

où u et v sont liés par la relation $u + v = m$.

Pour le démarrage de la récurrence, notons que :

$$P_m(2, \underline{u}) = \frac{\binom{u}{\ell}}{\binom{m}{\ell}} \times \frac{\binom{\ell}{2\ell - u}}{\binom{m}{\ell}} \quad ; \quad (87)$$

de sorte que

$$P_u(2, \underline{u}) = \frac{\binom{\ell}{2\ell - u}}{\binom{u}{\ell}} \quad . \quad (88)$$

D'autre part, en vertu de (83), on a :

$$Q_u(2,u) = P_u(2,\underline{u}) \quad . \quad (89)$$

Reprenons le second membre de (86). On peut montrer que le rapport des deux coefficients binomiaux (sous le signe [ ]) peut être approximé par :

$$\left(1-\frac{\ell}{m}\right)^{v} = \left(1-2^{-r}\right)^{v}.$$

Ainsi, la formule (86) de récurrence devient :

$$Q_m(k,m) \simeq \sum_{v}\left\{ \binom{\ell}{v} \beta^{v(k-1)} Q_{m-v}(k-1, m-v) \Big/ 0 \leqslant v \leqslant \ell \right\}, \quad (90)$$

où $\beta = 1-\frac{\ell}{m} = 1-2^{-r}$.

Ici encore, il sera intéressant de tabuler les probabilités $Q_m(k,m)$, en utilisant les formules (86) et (90) qui seront ainsi de plus comparées. Cette tabulation pourra se faire jusqu'à un niveau non négligeable, grâce à la puissance actuelle des ordinateurs.

IV.3.4. Se rendre compte que $k_0(r,n)$ est supérieur à $\mathcal{O}(K)$.

Rappelons que $k_0(r,n)$ a été défini dans la formule (67) ci-dessus. Il s'agira ici d'une illustration à propos de laquelle des calculs seront effectués qui permettront de répondre à l'énoncé du titre de ce paragraphe. Nous allons nous placer à une étape donnée où on suppose que, exactement $(m-i)$ points de Ω sont atteints. A partir de cette étape considérée comme l'état zéro du système, nous allons

comparer :

• le nombre d'étapes nécessaires pour que l'espérance mathématique du chargement complémentaire couvre les i derniers points;

• l'espérance mathématique du nombre d'étapes nécessaires pour atteindre les i derniers points.

Pour que le calcul soit le plus simple possible nous allons considérer trois cas correspondants à $i = 1, 2$ ou 3.

$\underline{i = 1}$

On établit par réccurrence que l'espérance mathématique du chargement complémentaire au bout de $(k+1)$ coups, s'écrit :

$$p\,(1 + q + \cdots + q^{j} + \cdots + q^{k}) \quad , \qquad (91)$$

où $p = \ell/m = 2^{-r}$ et $q = 1 - (\ell/m) = 1 - 2^{-r}$.

L'expression (91) peut se mettre sous la forme :

$$1 - q^{k+1} \quad , \qquad (92)$$

qui ne tend vers 1 que pour k tendant vers l'infini.

Considérons à présent

$$Pr\{K \geqslant k\} = \left[\frac{\binom{m-1}{\ell}}{\binom{m}{\ell}}\right]^{k-1} = q^{k-1} \quad . \qquad (93)$$

On obtient :

$$\mathcal{E}(K) = \sum_{k \geqslant 1} q^{k-1} = \frac{1}{1-q} = \frac{m}{\ell} = 2^r \quad , \quad (94)$$

qui est bien fini et ne dépend pas de m.

$\underline{i = 2}$

Le même principe du calcul par récurrence nous permet d'établir que l'espérance mathématique du changement complémentaire au bout de $(k+1)$ coups, est égale à

$$2 \wedge (1 + q + \ldots + q^k) = 2 (1 - q^{k+1}) \quad , \quad (95)$$

qui tend vers 2 pour k tendant vers l'infini.

Maintenant, en vertu de la formule d'inclusion et d'exclusion, on obtient :

$$Pr\{K \geqslant k\} = Pr\{K > k-1\} = 2 \left[\frac{\binom{m-1}{\ell}}{\binom{m}{\ell}}\right]^{k-1} - \left[\frac{\binom{m-2}{\ell}}{\binom{m}{\ell}}\right]^{k-1}$$

$$\simeq 2 q^{k-1} - q^{2(k-1)} \quad . \quad (96)$$

De sorte que

$$\mathcal{E}(K) = \sum_{k \geqslant 1} Pr\{K \geqslant k\} \simeq 2 \times \frac{1}{1-q} - \frac{1}{1-q^2}$$

$$= 2^r \times \frac{(3 \times 2^r - 2)}{(2 \times 2^r - 1)} \quad ; \quad (97)$$

ce qui, pour $r = 2$ donne $\mathcal{E}(K)$ compris entre 5 et 6.

$i = 3$

Le résultat correspondant à (95) devient :

$$3 + (1 + q + \cdots + q^{k}) = 3(1 - q^{k+1}) \qquad (98)$$

et celui, correspondant à (96) donne :

$$Pr\{K \geqslant k\} = Pr\{K > k-1\} =$$
$$3q^{k-1} - \binom{3}{2} q^{2(k-1)} + \binom{3}{3} q^{3(k-1)} \quad ; \quad (99)$$

ce qui donne, pour $\mathcal{E}(K) = \sum \{ Pr\{K \geqslant k\} / k \geqslant 1 \}$, la valeur suivante :

$$\frac{1}{2^{-r}} \left(3 - 3 \times \frac{1}{2 - 2^{-r}} + \frac{1}{3 - 2^{-r+1} - 2^{-r} + 2^{-2r}} \right) \quad , (100)$$

qui, pour $r = 2$, fournit la valeur 6,87... .

V. CONCLUSION

Au terme de notre étude nous espérons, à partir de l'analyse de [Simon & Dubois 1988], avoir présenté un regard nouveau sur les aspects formels et statistiques des problèmes de la complexité. Notre vision qui a un caractère ensembliste combinatoire et statistique [cf. § II], reste très liée à notre approche de la classification automatique des données qualitatives. Elle nous a permis de reprendre, avec une nouvelle formulation que nous considérons plus synthétique, des aspects traités dans la référence ci-dessus [cf. § III.1.1, § IV.2] sur le plan algorithme ou sur le plan calcul. Mais c'est, de façon

plus globale et plus variée, comprenant un guidage statistique pour faire «au mieux». D'autre part et surtout, nous avons mis au point de nouveaux algorithmes ou calculs, par rapport aux deux aspects fondamentaux que nous avons distingués : aspect évaluation et aspect reconnaissance [cf. § III.1. et § III.2. pour le cas observé, cf. § IV.1. et § IV.2. pour le cas aléatoire]. Nous avons pu, dans le cas observé, nous rendre compte de l'importance de la classification automatique pour la simplification de la complexité dans le cadre de l'évaluation (cf. § III.1.3.). D'autre part, nous avons introduit le parallélisme dans le cadre d'un algorithme de reconnaissance de la satisfiabilité (cf. § III.2.2.) dont il y a lieu d'étudier plus en détail la complexité.

Même si cette complexité doit devenir exponentielle, nous avons tout fait d'un point de vue statistique dans nos algorithmes, pour que, *a priori*, la conclusion (e.g. reconnaissance de la satisfiabilité) arrive le plus tôt possible.

Qu'il s'agisse d'un système réel observé de clauses, ou bien d'un système aléatoire, il est intéressant de constater que l'efficacité de la formule d'inclusion et d'exclusion est d'autant plus grande que le nombre de variables instanciées dans chacune des clauses, est grand (r grand, si le nombre de variables instanciées par clause est le même et égal à r) (cf. § III.1.2., § III.2.1. et § IV.2.3.).

Il faut savoir que la situation fournie par la réalisation d'un système aléatoire de clauses (représentés par des cylindres ponctuels), conformément à un des modèles de l'hypothèse d'absence de liaison, est celle, la plus défavorable en termes de complexité pour la recherche du nombre de solutions, ou de l'existence d'une solution. D'ailleurs, l'espérance mathématique du nombre de solutions est une portion de 2^m. Alors qu'en vérité, dans le cas de la donnée d'un système réel de clauses, représentés par des cylindres ponctuels, il y a des liens et des exclusions, qui impliquent quasiment toujours, une structure hiérarchique en classes et sous classes de dépendance statistique. La découverte d'une telle structure et son parcours ascendant permet dans la pratique de réduire très considérablement la complexité. Notre méthode de classification hiérarchique permet précisément une telle mise en évidence, où les classes et sous classes de dépendance sont repérés à partir de "nœuds significatifs" [Lerman 1981, 1991].

Références

[1] J.C. Simon et O. Dubois, "Number of solutions of satisfiability instances - Applications to knowledge bases", Rapport Laforia n°88/19, UA CNRS n° 1095 Univ. P. et M. Curie, Paris, Février 1988. Publié dans : International Journal of Pattern Recognition and Artificial Intelligence, Vol. 3, n° 1 (1989), 53-65.

[2] I.C. Lerman ; "Classification et analyse ordinale des données", Dunod, Paris (1981).

[3] I.C. Lerman ; "Foundations of the Likelihood Linkage Analysis (LLA) classification method", Applied Stochastic Models and Data Analysis, Vol 7, 63-76, Wiley (1991).

LISTE DES DERNIERES PUBLICATIONS INTERNES IRISA

PI 593 APPLICATION OF BELLEN'S PARALLEL METHOD TO ODE's WITH DISSIPATIVE RIGHT-HAND SIDE
Philippe CHARTIER
Juin 1991, 24 pages.

PI 594 PROGRAMMATION D'UN NOYAU UNIX EN GAMMA
Pascale LE CERTEN, Hector RUIZ BARRADAS
Juillet 1991, 48 pages.

PI 595 CALCULATING THE BUSY PERIOD DISTRIBUTION OF THE M/M/1 QUEUE
Louis-Marie LE NY, Gerardo RUBINO, Bruno SERICOLA
Juillet 1991, 11 pages.

PI 596 EFFICIENT CODE GENERATION FOR DISTRIBUTED MEMORY MACHINES
Françoise ANDRE, Olivier CHERON, Jean-Louis PAZAT, Henry THOMAS
Juillet 1991, 14 pages.

PI 597 KOAN : A SHARED VIRTUAL MEMORY FOR THE iPSC/2 HYPERCUBE
Zakaria LAHJOMRI, Thierry PRIOL
Juillet 1991, 32 pages.

PI 598 KOAN : A VERSATILE TOOL FOR PARALLELIZING REALISTIC RENDERING ALGORITHMS
Didier BADOUEL, Kadi BOUATOUCH, Zakaria LAHJOMRI, Thierry PRIOL
Juillet 1991, 28 pages.

PI 599 STATISTICAL ESTIMATION OF ROUNDOFF ERRORS AND CONDITION NUMBERS
Jocelyne ERHEL
Septembre 1991.

PI 600 NOMBRE DE SOLUTIONS ET SATISFIABILITE D'UN PROBLEME SAT ; UNE APPROCHE ENSEMBLISTE, COMBINATOIRE ET STATISTIQUE
Israël-César LERMAN
Septembre 1991.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique

ISSN 0249-6399